Squibs

From Annotator Agreement to Noise Models

Beata Beigman Klebanov*
Northwestern University

Eyal Beigman**
Northwestern University

This article discusses the transition from annotated data to a gold standard, that is, a subset that is sufficiently noise-free with high confidence. Unless appropriately reinterpreted, agreement coefficients do not indicate the quality of the data set as a benchmarking resource: High overall agreement is neither sufficient nor necessary to distill some amount of highly reliable data from the annotated material. A mathematical framework is developed that allows estimation of the noise level of the agreed subset of annotated data, which helps promote cautious benchmarking.

1. Introduction

By and large, the reason a computational linguist engages in an annotation project is to build a reliable data set for the eventual testing, and possibly training, of an algorithm performing the task. Hence, the crucial question regarding the annotated data set is whether it is good for benchmarking.

For classification tasks, the current practice is to infer this information from the value of an inter-annotator agreement coefficient such as the κ statistic (Cohen 1960; Siegel and Castellan 1988; Carletta 1996). If agreement is high, the whole of the data set is good for training and testing; the remaining disagreements are typically adjudicated by an expert (Snyder and Palmer 2004; Palmer, Kingsbury, and Gildea 2005; Girju, Badulescu, and Moldovan 2006) or through discussion (Litman, Hirschberg, and Swerts 2006), or, in case of more than two annotators, the majority label is chosen (Vieira and Poesio 2000).[1] There are some studies where cases of disagreement were removed from test data (Markert and Nissim 2002; Dagan, Glickman, and Magnini 2006). If agreement is low, the whole data set is discarded as unreliable. The threshold of acceptability seems to have stabilized around $\kappa = 0.67$ (Carletta 1996; Di Eugenio and Glass 2004).

There is little understanding, however, of exactly how and how well the value of κ reflects the quality of the data for benchmarking purposes. We develop a model of annotation generation that allows estimation of the level of noise in a specially constructed gold standard. A gold standard with a noise figure supports cautious benchmarking,

* Kellogg School of Management, Northwestern University, Evanston, IL, beata@northwestern.edu.
** Kellogg School of Management, Northwestern University, Evanston, IL, e-beigman@northwestern.edu.

1 In many studies, the procedure for handling disagreements is not clearly specified. For example, Gildea and Jurafsky (2002) mention a "consistency check"; in Lapata (2002), two annotators attained $\kappa = 0.78$ on 200 test instances, but it is not clear how cases of disagreements were settled.

Submission received: 30 June 2008; revised submission received: 3 December 2008; accepted for publication: 26 January 2009.

by requiring that the performance of an algorithm be better than baseline by more than that which can be attributed to noise. Articulating an annotation generation model also allows us to shed light on the information κ can contribute to benchmarking.

2. Annotation Noise

We are interested in finding out which parts of the annotated data are sufficiently reliable. This question presupposes a division of instances into two types: reliable and unreliable, or, as we shall call them, **easy** and **hard**, under the assumption that items that are easy are reliably annotated, whereas items that are hard display confusion and disagreement. The plausibility of separation into easy and hard instances is supported by researchers conducting annotation projects: "With many judgments that characterize natural language, one would expect that there are clear cases as well as borderline cases that are more difficult to judge" (Wiebe, Wilson, and Cardie 2005, page 200).

This suggests a model of annotation generation with latent variables for types, thus, for every instance i, there is a variable l_i with values E (easy) and H (hard). Let n be the number of instances, k the number of annotators, and X_{ij} the classification of the ith instance by the jth annotator. An annotation generation model assigns a functional form to the joint distribution conditioned on the latent variable $\mathbb{P}(X_{i1}, \ldots, X_{ik}|l_i)$. Similar models have been studied in biometrics (Aickin 1990; Hui and Zhou 1998; Albert, McShane, and Shih 2001; Albert and Dodd 2004). The main assumption is that, conditioned on the type, annotators agree on easy instances and independently flip a coin on hard ones. The joint distribution satisfies:

$$\mathbb{P}(X_{i1} = \ldots = X_{ik}|l_i = E) = 1; \quad \mathbb{P}(X_{i1} = b_1, \ldots, X_{ik} = b_k|l_i = H) = \prod_{j=1}^{k} \mathbb{P}(X_{ij} = b_j|l_i = H)$$

We want to take only easy instances into the gold standard, so that it contains only settled, trustworthy judgments.[2] The problem is that the fact of being easy or hard is not directly observable, but has to be inferred from the observed annotations. In particular, some of the observed agreements will in fact be hard instances, since coin-flips could occasionally come out all-heads or all-tails. Our objective is to estimate, with a given degree of confidence (α), the proportion γ of hard instances in the agreed annotations, based on the number of observed disagreements. The value of γ is the level of **annotation noise** in the gold standard comprising agreed annotations.

Let p be the probability that the annotators agree on a hard instance in a binary classification task:

$$p = \mathbb{P}(X_{i1} = \ldots = X_{ik}|l_i = H) = \prod_{j=1}^{k} \mathbb{P}(X_{ij} = 0|l_i = H) + \prod_{j=1}^{k} \mathbb{P}(X_{ij} = 1|l_i = H)$$

Denote by A_d the event that there are d disagreed instances; these are hard, and are assumed to be labeled by coin-flips. Let B_h be the event that there are overall h hard

2 On the status of hard instances, see Section 5.1.

instances; some of these may be unobserved as they surface as random agreements. We note that $\mathbb{P}(A_d|B_h) = \binom{h}{d} \cdot (1-p)^d \cdot p^{h-d}$ for $d \leq h$, hence:

$$\mathbb{P}(B_h|A_d) = \frac{\mathbb{P}(A_d \cap B_h)}{\mathbb{P}(A_d)} = \frac{\mathbb{P}(A_d|B_h) \cdot \mathbb{P}(B_h)}{\sum_{i=d}^{n} \mathbb{P}(A_d|B_i) \cdot \mathbb{P}(B_i)} = \frac{\binom{h}{d} \cdot p^{h-d} \cdot \mathbb{P}(B_h)}{\sum_{i=d}^{n} \binom{i}{d} \cdot p^{i-d} \cdot \mathbb{P}(B_i)}$$

Let X be a random variable designating the number of coin-flips. It follows that

$$\mathbb{P}(X > t|A_d) = \frac{\sum_{i=t+1}^{n} \binom{i}{d} \cdot p^{i-d} \cdot \mathbb{P}(B_i)}{\sum_{i=d}^{n} \binom{i}{d} \cdot p^{i-d} \cdot \mathbb{P}(B_i)} \quad (1)$$

Let t_0 be the smallest integer for which $\mathbb{P}(X > t_0|A_d) < 1 - \alpha$. Given d observed disagreements, we estimate the noise level of the agreed subset of the annotations as at most $\gamma = \frac{t_0 - d}{n - d}$, with confidence α.

3. Relation to κ Statistic

3.1 The Case of High κ with Two Annotators

Suppose 1,000 instances have been annotated by two people, such that 900 are instances of agreement. Both in the 900 agreed instances and in the 100 disagreed ones, the categories were estimated to be equiprobable for both annotators.[3] In this case $p = 0.5$, $\kappa = 0.8$,[4] which is usually taken to be an indicator of sufficiently agreeable guidelines, and, by implication, of a high quality data set. Our candidate gold standard is the 900 instances of agreement. What is its 95% confidence noise rate? We find, using our model, that with more than 5% probability up to 125 agreements are due to coin-flipping, hence $\gamma = 13.8\%$.[5] This scenario is not hypothetical. In Poesio and Vieira (1998) Experiment 1, the classification of definite descriptions into Anaphoric-or-Associative versus Unfamiliar has $n = 992$, $d = 121$, $p = 0.47$, which, with 95% confidence, yields $\gamma = 15\%$.

Let us reverse the question: For a two-annotator project with 1,000 instances, how many disagreements could we tolerate, so that the agreed part is 95% noise-free with 95% confidence? Only 33 disagreements, corresponding to $\kappa = 0.93$. In practice, this means that a two-annotator project of this size is unlikely to produce a high-quality gold standard, the high κ notwithstanding.

3.2 The Case of Low κ with Five Annotators

Suppose now 1,000 instances are annotated by five people, with 660 agreements. With categories equiprobable in both hard and easy instances, p = 0.0625. The exact value of κ depends on the distribution of votes in the 340 disagreed cases, from $\kappa = 0.73$ when all disagreements are split 4-to-1, to $\kappa = 0.52$ when all disagreements are split 3-to-2. Assuming disagreements are coin-flips, the most likely measurement would be about $\kappa = 0.637$, where the 340 observed coin-flips yielded the most likely pattern.[6] This value of κ is considered low, yet the 660 agreed items make a gold standard within the

3 We estimate $\mathbb{P}(X_{ij} = 1|l_i = H)$ by the proportion of disagreed instances that annotator j put in category 1.
4 For calculating κ, we use the version shown in Equation (2).
5 In all our calculations $\mathbb{P}(B_1) = \ldots = \mathbb{P}(B_n)$, that is, a priori, any number of hard instances is equiprobable.
6 That is, there are twice as many 3-to-2 cases than 4-to-1, corresponding to $\binom{5}{3}$ as opposed to $\binom{5}{4}$.

noise rate of $\gamma = 5\%$ with 95% confidence, according to our model. Hence it is possible for the overall annotation to have low-ish κ, but the agreement of all five annotators, if observed sufficiently frequently, is reliable, and can be used to build a clean gold standard.

3.3 Interpreting the κ Statistic in the Annotation Generation Model

The κ statistic is defined as $\kappa = \frac{P_A - P_E}{1 - P_E}$ where P_A is the observed agreement and P_E is the agreement expected by chance, calculated from the marginals. We use the Siegel and Castellan (1988) version, referred to as K in Artstein and Poesio (2008):

$$P_E = \sum_{j=1}^{m} p_j^2; \;\; p_j = \frac{\sum_{i=1}^{n} a_{ij}}{nk}; \;\; P_A = \frac{1}{n}\sum_{i=1}^{n} P_{A_i}; \;\; P_{A_i} = \frac{\sum_{j=1}^{m} \binom{a_{ij}}{2}}{\binom{k}{2}} \tag{2}$$

where n is the number of items; m is the number of categories; k is the number of annotators; and a_{ij} is the number of annotators who assigned the ith item to the jth category.

Suppose there are h hard instances and e easy ones, and $m = 2$. Suppose further that all annotators flip the same coin on hard instances, and that the distribution of the categories in easy and hard instances is the same and is given by $q_1, \ldots, q_m$. Then the probability for chance agreement between two annotators is $q = \sum_{j=1}^{m} q_j^2$, of which P_E is an estimator. Agreement on a particular instance P_{A_i} is measured by the proportion of agreeing pairs of annotators out of all such pairs, and P_A is an estimator of the expected agreement across all instances. Our model assumes perfect agreement on easy instances and agreement with probability q on hard ones, so we expect to see $e + q \cdot h$ agreed instances, hence P_A is an estimator of $\frac{e+qh}{e+h}$. Putting these together, $\kappa = \frac{P_A - P_E}{1 - P_E}$ is an estimator of $\frac{\frac{e+qh}{e+h} - q}{1-q} = \frac{e}{e+h}$, the proportion of easy instances.[7] In fact, Aickin (1990) shows that κ is very close to this ratio when the marginal distribution over the categories is uniform, with a more substantial divergence for skewed category distributions.[8]

The correspondence between κ and the proportion of easy instances makes it clear why κ is not a sufficient indicator of data quality for benchmarking. For when $\kappa = 0.8$, 20% of the data are hard cases. Using all data, especially for testing, is thus potentially hazardous, and the crucial question is: Can we zero in on the easy instances effectively, without admitting much noise? This is exactly the question answered by the model.

When the distribution of categories is the same in easy and hard instances and uniform, κ can be used to address this question as well. Recall that in the two-annotator case in Section 3.1, $\kappa = 0.8$, that is, 80% of instances are estimated to be easy. Because easy cases are a subset of agreed ones in our model, 800 of the agreed 900 instances are easy, giving an estimate of 11% noise in the gold standard. Requiring 95% confidence in noise estimation, we found $\gamma = 13.8\%$, using our model. Similarly, in the five-annotator

7 The proportion of easy cases is positive, whereas the estimator κ can be negative with non-negligible probability when $e = O(\sqrt{h})$.

8 In Aickin (1990), category distribution on easy cases is derived from that in the hard cases. The closer the categories are to uniform distribution in the hard cases, the closer their distribution in hard cases is to that in easy cases. For example, if the categories are distributed uniformly in hard cases, they are also so distributed in the easy ones. If the categories are distributed $(\frac{1}{3}, \frac{2}{3})$ in the hard cases, they are distributed $(\frac{1}{5}, \frac{4}{5})$ in the easy cases. For this reason, in Aickin's model, it is not possible to distinguish between category imbalance (many more 0s than 1s) and differences in category distributions in easy and hard cases. His simulations show that in cases of category imbalance (which imply, in his model, differences in category distributions in easy and hard cases), κ tends to underestimate the proportion of easy instances.

scenario in Section 3.2, $\kappa = 0.637$ tells us that about 637 out of 1,000 instances are easy; they are captured quite precisely by the 660 agreements, yielding a noise estimate of 3.5%, again somewhat lower than the high confidence one we gave using the model.

4. Training and Testing in the Presence of Annotation Noise

We discuss two uses of a gold standard within the benchmarking enterprise. The data could be used for testing, and, if there is enough of it and after an appropriate partition, for training as well. We consider each case separately in the following sections.

4.1 Testing with Annotation Noise

The two questions one wants to answer using the data are: How well does an algorithm capture the phenomenon? For any two algorithms, which one is better? Consider the algorithm comparison situation. Suppose we have a gold standard with L items of which up to R are noise ($\gamma = \frac{R}{L}$). Two algorithms might differ in performance on the easy cases, the hard ones, or both. Because we cannot distinguish between easy and hard instances in the gold standard, we are unable to attribute the difference in performance correctly. Moreover, as the annotations of the hard instances are random coin-flips, there is an expected difference in performance that is a result of pure chance.

Suppose two algorithms perform equally well on easy instances; their performance on the hard ones is as good as agreement-by-coin-flipping would allow. Thus, the difference in the number of "correct" answers on hard instances for algorithms A and B is a random variable S satisfying $S = \sum_{i=1}^{R} X_i$ where $X_1, \ldots, X_R$ are independent and identically distributed random variables which obtain values -1 (A "right", B "wrong") and 1 (A "wrong", B "right") with probability $\frac{1}{4}$ and 0 with probability $\frac{1}{2}$, thus $\mu_S = 0$; $\sigma_S = \sqrt{\frac{R}{2}}$. By Chebyshev's inequality $Pr(|S| > k\sigma) \leq \frac{1}{k^2}$: that is, the chance difference between the algorithms will be within 4.5σ with 95% probability.[9] In our example, L = 900 and R = 125, hence a difference of up to 35 "correct" answers (3.9% of the gold standard) can be attributed to chance.[10]

This example shows that even if getting a clean data set is not feasible, it is important to report the noise rate of the data set that has been produced. This would allow calibrating the benchmarking procedure by requiring the difference between the two competing algorithms to be larger than the chance difference scale.

Some perils of testing on noisy data were discussed in a recent article in this journal by Reidsma and Carletta (2008). They showed that a machine-learning classifier is sensitive to the type of noise in the data. Specifically, if the noise is in the form of category over-use (an annotator disproportionately favors a certain category), when algorithm performance is measured against the noisy data, accuracy estimates are often inflated relative to performance on the real data, uncorrupted by noise (see Figure 3(b) therein). This is because "when the observed data is used to test performance, some of

9 For large R, normal approximation can be used with the tighter 2σ bound for 95% confidence.

10 We note that because the difference attributable to coin-flipping is $O(\sqrt{\frac{\gamma}{L}})$, and assuming noise rate is constant, the scale of chance difference diminishes with larger data sets (see also footnote 9). The issue is more important when dealing with small-to-moderate data sets. However, even for a 130K test set (Sections 22–24 of the Wall Street Journal corpus, standardly used as a test set in POS-tagging benchmarks), it is useful to know the estimated noise rate, as it is not clear that all reported improvements in performance would come out significant. For example, Shen, Satta, and Joshi (2007) summarize performance of five previously published and three newly reported algorithms, all between 97.10% and 97.33%.

the samples match not because the classifier gets the label right, but because it overuses the same label as the human coder" (Reidsma and Carletta 2008, page 232). On the other hand, if disagreements are random classification noise (the label of any instance can be flipped with a certain probability), a performance estimate based on observed data would often be lower than performance on the real data, because the noise that corrupted it was ignored by the classifier (see Figure 2(d) therein).

Reidsma and Carletta (2008) suggest that the community develops methods to investigate the patterns of disagreements between annotators to gain insight into the potential of incorrect performance estimation. Although we agree on the general point that human agreements and disagreements should bear directly on the practice of estimating the performance of an algorithm, we focus on improving the quality of performance estimation. We suggest (1) mitigating the effect of annotation noise on performance estimation by using the least noisy part of the data set for testing, that is, a gold standard with agreed items; (2) providing an estimate of the level of noise in the gold standard, which can be used to gauge the divergence between the estimate of performance using the gold standard from the real performance figure on the easy instances (i.e., on noise-free data), similarly to the algorithm comparison scenario provided herein.

4.2 Learning with Annotation Noise

The problem with noise in the training data is the potential for misclassification of easy instances in the test data as a result of hard instances in the training data, the problem we call **hard case bias**.

Learning in the presence of noise is an active research area in machine learning. However, annotation noise is different from existing well-understood noise models. Specifically, random classification noise, where each instance has the same probability of having its label flipped, is known to be tolerable in supervised learning (Blum et al. 1996; Cohen 1997; Reidsma and Carletta 2008). In annotation noise, coin-flipping is confined to hard instances, which should not be assumed to be uniformly distributed across the feature space. Indeed, there is reason to believe that they form clusters; certain feature combinations tend to give rise to hard instances. The finding reported by Reidsma and op den Akker (2008) that a classifier trained on data from one annotator tended to agree much better with test data from the same annotator than with that of another annotator exemplifies a situation where observed hard cases (i.e., cases where the annotators disagree) constitute a pattern in the feature space that a classifier picks up.

In a separate article, we establish a number of properties of learning under annotation noise (Beigman and Beigman Klebanov 2009). We show that the 0-1 loss model may be vulnerable to annotation noise for small data sets, but becomes increasingly robust the larger the data set, with worst-case hard case bias of $\theta(\frac{1}{\sqrt{n}})$. We also show that learning with the popular voted-perceptron algorithm (Freund and Schapire 1999) could suffer a constant rate of hard case bias irrespective of the size of the data set.

5. Discussion

5.1 The Status of Hard Instances

We suggested that only the easy instances should be taken into the gold standard. This is not to say that hard cases should be eliminated from the researcher's attention; we merely argue that they should not be used for testing algorithms for benchmarking

purposes. Hard cases are interesting for theory development, because this is where the theory might have a difficulty, but they do not allow for a fair comparison, as their correct label cannot be determined under the current theory. The agreed data embodies the well-articulated parts of the theory, which are ready for deployment as a gold standard for machine learning. Once the theory is improved to a stage where some of the previously hard cases receive an unproblematic treatment, those items can be added to the data set, which can make the task more challenging for the machine. Linguistic theories-in-the-making can have limited coverage; they do not immediately attain the status of medical conditions, for example, where there presumably exists a true label even for the hardest-to-diagnose cases.[11]

5.2 Plausibility of the Model

Beyond the separation into easy and hard instances, our model prescribes certain annotator behavior for each type. In our work on metaphor, we observed that certain metaphor markups were retracted by their authors, when asked after 4–8 weeks to revisit the annotations (Beigman Klebanov, Beigman, and Diermeier 2008). These were apparently hard cases, with people resolving their doubts inconsistently on the two occasions; coin-flipping is a reasonable first-cut model for such cases. The model also accommodates category over-use bias (Di Eugenio and Glass 2004; Artstein and Poesio 2008; Reidsma and Carletta 2008), as $\mathbb{P}(X_{ij} = b_j | l_i = H)$ may vary across annotators.

Still, this model is clearly a simplification. For example, it is possible that there is more than one degree of hardness, and annotator behavior changes accordingly. Another extension is modeling imperfect annotators, allowed to commit random errors on easy cases; this extension would be needed if a large number of annotators is used.

Such extensions, as well as methods for estimating these more complex models, should clearly be put on the community's research agenda. The main contribution of the simple model is in outlining the trajectory from agreement to gold standard with a noise estimate, and indicating the potential benefit of the latter to data utilization (low overall agreement does not preclude the existence of a reliable subset) and to prudent benchmarking. Furthermore, the simple model helps us improve the understanding of the information provided by the κ statistic, and to appreciate its limitations. It also allows us to see the benefit of adding annotators, as discussed in the next section.

5.3 Adding Annotators

If we want the test data to be able to detect small advances in machines' handling of the task, we need to produce gold standards with low noise levels. The level of noise in agreed data depends on two parameters: (a) the number of agreed items, and (b) the probability of chance agreement between annotators. Although the first is not under the researcher's control once the data set is chosen, the second is, by changing the number of annotators. Obviously, the more annotators are required to agree, the lower p will be, and the smaller the number of agreements that can be attributed to coin-flipping. If indeed 800 out of 1,000 items are easy, agreement between two annotators can only detect them with up to 13.8% noise. Adding a third annotator means $p = 0.25$.

11 As one of the anonymous reviewers pointed out, some medical conditions, such as autism, are also only partially understood.

We are most likely to observe 850 agreed instances, which would not contain more than 7.7% noise, with 95% confidence. Effectively, we got rid of about half the random agreements.

Acknowledgments
We thank Eli Shamir and Bei Yu for reading earlier drafts of this article, as well as the editor and the anonymous reviewers for comments that helped us improve the article significantly.

References

Aickin, Mikel. 1990. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*, 46(2):293–302.

Albert, Paul and Lori Dodd. 2004. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435.

Albert, Paul, Lisa McShane, and Joanna Shih. 2001. Latent class modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57(2):610–619.

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

Beigman, Eyal and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, Singapore.

Beigman Klebanov, Beata, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *COLING 2008 Workshop on Human Judgments in Computational Linguistics*, pages 2–7, Manchester.

Blum, Avrim, Alan Frieze, Ravi Kannan, and Santosh Vempala. 1996. A polynomial-time algorithm for learning noisy linear threshold functions. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, pages 330–338, Burlington, VT.

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

Cohen, Edith. 1997. Learning noisy perceptrons by a perceptron in polynomial time. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, pages 514–523, Miami Beach, FL.

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

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *The PASCAL Recognising Textual Entailment Challenge*, Springer, Berlin, pages 177–190.

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

Freund, Y. and R. E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Girju, Roxana, Adriana Badulescu, and Dan Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Hui, Siu and Xiao Zhou. 1998. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7(4):354–370.

Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Litman, Diane, Julia Hirschberg, and Marc Swerts. 2006. Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics*, 32(3):417–438.

Markert, Katja and Malvina Nissim. 2002. Metonymy resolution as a classification task. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 204–213, Philadelphia, PA.

Palmer, Martha, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

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

Reidsma, Dennis and Rieks op den Akker. 2008. Exploiting subjective annotations. In *COLING 2008 Workshop on Human Judgments in Computational Linguistics*, pages 8–16, Manchester.

Shen, Libin, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 760–767, Prague.

Siegel, Sidney and N. John Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

Snyder, Benjamin and Martha Palmer. 2004. The English all-words task. In *Senseval-3: 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona.

Vieira, Renata and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Wiebe, Janyce, Teresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.