# To Annotate More Accurately or to Annotate More

**Dmitriy Dligach**
Department of Computer Science
University of Colorado at Boulder
Dmitriy.Dligach@colorado.edu

**Rodney D. Nielsen**
The Center for Computational Language
and Education Research
University of Colorado at Boulder
Rodney.Nielsen@colorado.edu

**Martha Palmer**
Department of Linguistics
Department of Computer Science
University of Colorado at Boulder
Martha.Palmer@colorado.edu

## Abstract

The common accepted wisdom is that blind double annotation followed by adjudication of disagreements is necessary to create training and test corpora that result in the best possible performance. We provide evidence that this is unlikely to be the case. Rather, the greatest value for your annotation dollar lies in single annotating more data.

## 1 Introduction

In recent years, supervised learning has become the dominant paradigm in Natural Language Processing (NLP), thus making the creation of hand-annotated corpora a critically important task. A corpus where each instance is annotated by a single tagger unavoidably contains errors. To improve the quality of the data, an annotation project may choose to annotate each instance twice and adjudicate the disagreements, thus producing the (largely) error-free gold standard. For example, OntoNotes (Hovy et al., 2006), a large-scale annotation project, chose this option.

However, given a virtually unlimited supply of unlabeled data and limited funding – a typical set of constraints in NLP – an annotation project must always face the realization that for the cost of double annotation, more than twice as much data can be *single* annotated. The philosophy behind this alternative says that modern machine learning algorithms can still generalize well in the presence of noise, especially when given larger amounts of training data.

Currently, the commonly accepted wisdom sides with the view that says that blind double annotation followed by adjudication of disagreements is necessary to create annotated corpora that leads to the best possible performance. We provide empirical evidence that this is unlikely to be the case. Rather, the greatest value for your annotation dollar lies in single annotating more data. There may, however, be other considerations that still argue in favor of double annotation.

In this paper, we also consider the arguments of Beigman and Klebanov (2009), who suggest that data should be multiply annotated and then filtered to discard all of the examples where the annotators do not have perfect agreement. We provide evidence that single annotating more data for the same cost is likely to result in better system performance.

This paper proceeds as follows: first, we outline our evaluation framework in Section 2. Next, we compare the single annotation and adjudication scenarios in Section 3. Then, we compare the annotation scenario of Beigman and Klebanov (2009) with the single annotation scenario in Section 4. After that, we discuss the results and future work in section 5. Finally, we draw the conclusion in Section 6.

## 2 Evaluation

### 2.1 Data

For evaluation we utilize the word sense data annotated by the OntoNotes project. The OntoNotes data was chosen because it utilizes full double-blind annotation by human annotators and the disagreements are adjudicated by a third (more expe-

*Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*, pages 64–72,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

rienced) annotator. This allows us to

- Evaluate single annotation results by using the labels assigned by the first tagger
- Evaluate double annotation results by using the labels assigned by the second tagger
- Evaluate adjudication results by using the labels assigned by the the adjudicator to the instances where the two annotators disagreed
- Measure the performance under various scenarios against the double annotated and adjudicated gold standard data

We selected the 215 most frequent verbs in the OntoNotes data. To make the size of the dataset more manageable, we randomly selected 500 examples of each of the 15 most frequent verbs. For the remaining 200 verbs, we utilized all the annotated examples. The resulting dataset contained 66,228 instances of the 215 most frequent verbs. Table 1 shows various important characteristics of this dataset averaged across the 215 verbs.

| | |
|---|---|
| Inter-tagger agreement | 86% |
| Annotator1-gold standard agreement | 93% |
| Share of the most frequent sense | 70% |
| Number of classes (senses) per verb | 4.74 |

Table 1: Data used in evaluation at a glance

### 2.2 Cost of Annotation

Because for this set of experiments we care primarily about the cost effectiveness of the annotation dollars, we need to know how much it costs to blind annotate instances and how much it costs to adjudicate disagreements in instances. There is an upfront cost associated with any annotation effort to organize the project, design an annotation scheme, set up the environment, create annotation guidelines, hire and train the annotators, etc. We will assume, for the sake of this paper, that this cost is fixed and is the same regardless of whether the data is single annotated or the data is double annotated and disagreements adjudicated.

In this paper, we focus on a scenario where there is essentially no difference in cost to collect additional data to be annotated, as is often the case (e.g., there is virtually no additional cost to download 2.5 versus 1.0 million words of text from the web). However, this is not always the case (e.g., collecting speech can be costly).

To calculate a cost per annotated instance for blind annotation, we take the total expenses associated with the annotators in this group less training costs and any costs not directly associated with annotation and divide by the total number of blind instance annotations. This value, \$0.0833, is the per instance cost used for single annotation. We calculated the cost for adjudicating instances similarly, based on the expenses associated with the adjudication group. The adjudication cost is an additional \$0.1000 per instance adjudicated. The per instance cost for double blind, adjudicated data is then computed as double the cost for single annotation plus the per instance cost of adjudication multiplied by the percent of disagreement, 14%, which is \$0.1805.

We leave an analysis of the extent to which the up front costs are truly fixed and whether they can be altered to result in more value for the dollar to future work.

### 2.3 Automatic Word Sense Disambiguation

For the experiments we conduct in this study, we needed a word sense disambiguation (WSD) system. Our WSD system is modeled after the state-of-the-art verb WSD system described in (Dligach and Palmer, 2008). We will briefly outline it here.

We view WSD as a supervised learning problem. Each instance of the target verb is represented as a vector of binary features that indicate the presence (or absence) of the corresponding features in the neighborhood of the target verb. We utilize all of the linguistic features that were shown to be useful for disambiguating verb senses in (Chen et al., 2007).

To extract the **lexical features** we POS-tag the sentence containing the target verb and the two surrounding sentences using MXPost software (Ratnaparkhi, 1998). All open class words (nouns, verbs, adjectives, and adverbs) in these sentences are included in our feature set. In addition to that, we use as features two words on each side of the target verb as well as their POS tags.

To extract the **syntactic features** we parse the sentence containing the target verb with Bikel's constituency parser and utilize a set of rules to identify the features in Table 2.

Our **semantic features** represent the semantic classes of the target verb's syntactic arguments

| Feature | Explanation |
|---|---|
| Subject and object | - Presence of subject and object<br>- Head word of subject and object NPs<br>- POS tag of the head word of subject and object NPs |
| Voice | - Passive or Active |
| PP adjunct | - Presence of PP adjunct<br>- Preposition word<br>- Head word of the preposition's NP argument |
| Subordinate clause | - Presence of subordinate clause |
| Path | - Parse tree path from target verb to neighboring words<br>- Parse tree path from target verb to subject and object<br>- Parse tree path from target verb to subordinate clause |
| Subcat frame | - Phrase structure rule expanding the target verb's parent node in parse tree |

Table 2: Syntactic features

such as subject and object. The semantic classes are approximated as

- WordNet (Fellbaum, 1998) hypernyms
- NE tags derived from the output of IdentiFinder (Bikel et al., 1999)
- Dynamic dependency neighbors (Dligach and Palmer, 2008), which are extracted in an unsupervised way from a dependency-parsed corpus

Our WSD system uses the Libsvm software package (Chang and Lin, 2001) for classification. We accepted the default options (C = 1 and linear kernel) when training our classifiers. As is the case with most WSD systems, we train a separate model per verb.

## 3 Experiment One

The results of experiment one show that in these circumstances, better performance is achieved by single annotating more data than by deploing resources towards ensuring that the data is annotated more accurately through an adjudication process.

### 3.1 Experimental Design

We conduct a number of experiments to compare the effect of single annotated versus adjudicated data on the accuracy of a state of the art WSD system. Since OntoNotes does not have a specified test set, for each word, we used repeated random partitioning of the data with 10 trials and 10% into the test set and the remaining 90% comprising the training set.

We then train an SVM classifier on varying fractions of the data, based on the number of examples that could be annotated per dollar. Specifically, in increments of \$1.00, we calculate the number of examples that can be single annotated and the number that can be double blind annotated and adjudicated with that amount of money.

The number of examples computed for single annotation is selected at random from the training data. Then the adjudicated examples are selected at random from this subset. Selecting from the same subset of data approaches pair statistical testing and results in a more accurate statistical comparison of the models produced.

Classifiers are trained on this data using the labels from the first round of annotation as the single annotation labels and the final adjudicated labels for the smaller subset. This procedure is repeated ten times and the average results are reported.

For a given verb, each classifier created throughout this process is tested on the same double annotated and adjudicated held-out test set.

### 3.2 Results

Figure 1 shows a plot of the accuracy of the classifiers relative to the annotation investment for a typical verb, *to call*. As can be seen, the accuracy is always higher when training on the larger amount of single annotated data than when training on the amount of adjudicated data that had the equivalent cost of annotation.

Figures 2 and 3 present results averaged over all 215 verbs in the dataset. First, figure 2 shows the average accuracy over all verbs by amount invested. These accuracy curves are not smooth be-

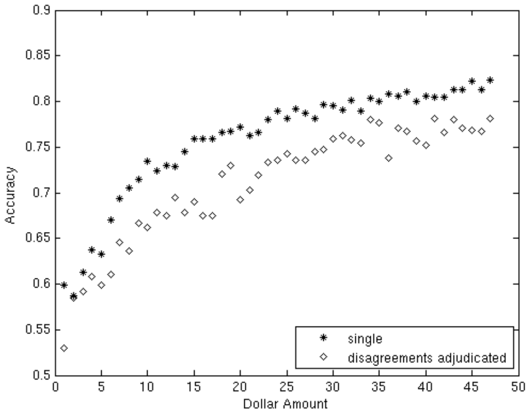

Figure 1: Performance of single annotated vs. adjudicated data by amount invested for *to call*

cause the verbs all have a different number of total instances. At various annotation cost values, all of the instances of one or more verbs will have been annotated. Hence, the accuracy values might jump or drop by a larger amount than seen elsewhere in the graph.

Toward the higher dollar amounts the curve is dominated by fewer and fewer verbs. We only display the dollar investments of up to $60 due to the fact that only five verbs have more than $60's worth of instances in the training set.

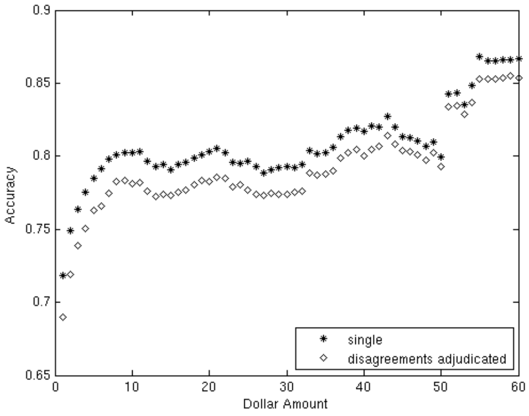

Figure 2: Average performance of single annotated vs. adjudicated data by amount invested

The average difference in accuracy for Figure 2 across all amounts of investment is 1.64%.

Figure 3 presents the average accuracy relative to the percent of the total cost to single annotate all of the instances for a verb. The accuracy at a given percent of total investment was interpolated for each verb using linear interpolation and then averaged over all of the verbs.

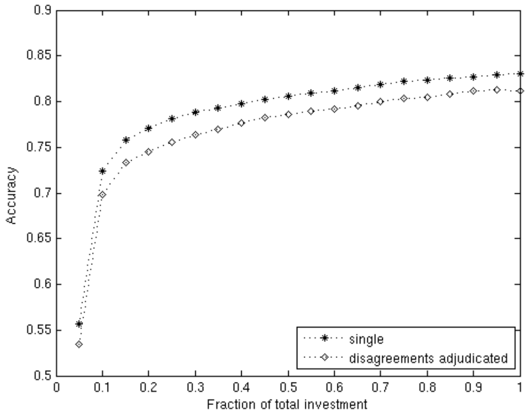

Figure 3: Average performance of single annotated vs. adjudicated data by fraction of total investment

The average difference in accuracy for Figure 3 across each percent of investment is 2.10%.

Figure 4 presents essentially the same information as Figure 2, but as a reduction in error rate for single annotation relative to full adjudication.

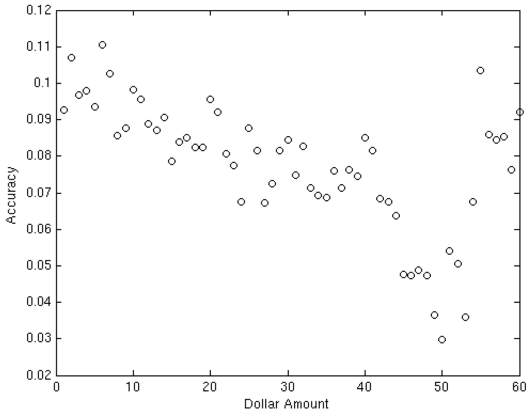

Figure 4: Reduction in error rate from adjudication to single annotation scenario based on results in Figure 2

The relative reduction in error rate averaged over all investment amounts in Figure 2 is 7.77%.

Figure 5 presents the information in Figure 3 as a reduction in error rate for single annotation relative to full adjudication.

The average relative reduction in error rate over the fractions of total investment in Figure 5 is 9.32%.

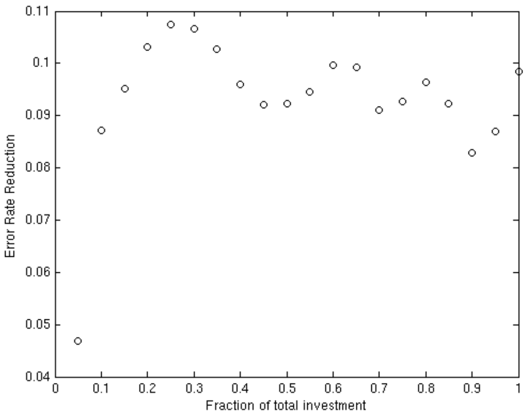

Figure 5: Reduction in error rate from adjudication to single annotation scenario based on results in Figure 3

### 3.3 Discussion

First, it is worth noting that, when the amount of annotated data is the *same* for both scenarios, adjudicated data leads to slightly better performance than single annotated data. For example, consider Figure 3. The accuracy at 100% of the total investment for the double annotation and adjudication scenario is 81.13%. The same number of examples can be *single* annotated for 0.0833 / 0.1805 = 0.4615 of this dollar investment (using the costs from Section 2.2). The system trained on that amount of single annotated data shows a lower accuracy, 80.21%. Thus, in this case, the adjudication scenario brings about a performance improvement of about 1%.

However, the main thesis of this paper is that instead of double annotating and adjudicating, it is often better to single annotate more data because it is a more cost-effective way to achieve a higher performance. The results of our experiments support this thesis. At every dollar amount invested, our supervised WSD system performs better when trained on single annotated data comparing to double annotated and adjudicated data.

The maximum annotation investment amount for each verb is the cost of single annotating all of its instances. When the system is trained on the amount of double annotated data possible at this investment, its accuracy is 81.13% (Figure 3). When trained on single annotated data, the system attains the same accuracy much earlier, at approximately 60% of the total investment. When trained on the entire available single annotated data, the system reaches an accuracy of 82.99%, nearly a 10% relative reduction in error rate over the same system trained on the adjudicated data obtained for the same cost.

Averaged over the 215 verbs, the single annotation scenario outperformed adjudication at every dollar amount investigated.

## 4 Experiment Two

In this experiment, we consider the arguments of Beigman and Klebanov (2009). They suggest that data should be at least double annotated and then filtered to discard all of the examples where there were any annotator disagreements.

The main points of their argument are as follows. They first consider the data to be dividable into two types, *easy* (to annotate) *cases* and *hard cases*. Then they correctly note that some annotators could have a systematic bias (i.e., could favor one label over others in certain types of hard cases), which would in turn bias the learning of the classifier. They show that it is theoretically possible that a band of misclassified hard cases running parallel to the true separating hyperplane could mistakenly shift the decision boundary past up to $\sqrt{N}$ easy cases.

We suggest that it is extremely unlikely that a consequential number of easy cases would exist nearer to the class boundary than the hard cases. The hard cases are in fact generally considered to define the separating hyperplane.

In this experiment, our goal is to determine how the accuracy of classifiers trained on data labeled according to Beigman and Klebanov's *discard disagreements* strategy compares empirically to the accuracy resulting from single annotated data. As in the previous experiment, this analysis is performed relative to the investment in the annotation effort.

### 4.1 Experimental Design

We follow essentially the same experimental design described in section 3.1, using the same state of the art verb WSD system. We conduct a number of experiments to compare the effect of single annotated versus double annotated data. We utilized the same training and test sets as the previous experiment and similarly trained an SVM on fractions of the data representing increments of $1.00 investments.

As before, the number of examples designated for single annotation is selected at random from the training data and half of that subset is selected as the training set for the double annotated data. Again, selecting from the same subset of data results in a more accurate statistical comparison of the models produced.

Classifiers for each annotation scenario are trained on the labels from the first round of annotation, but examples where the second annotator disagreed are thrown out of the double annotated data. This results in slightly less than half as much data in the double annotation scenario based on the disagreement rate. Again, the procedure is repeated ten times and the average results are reported.

For a given verb, each classifier created throughout this process is tested on the same double annotated and adjudicated held-out test set.

### 4.2 Results

Figure 6 shows a plot of the accuracy of the classifiers relative to the annotation investment for a typical verb, *to call*. As can be seen, the accuracy for a specific investment performing single annotation is always higher than it is for the same investment in double annotated data.

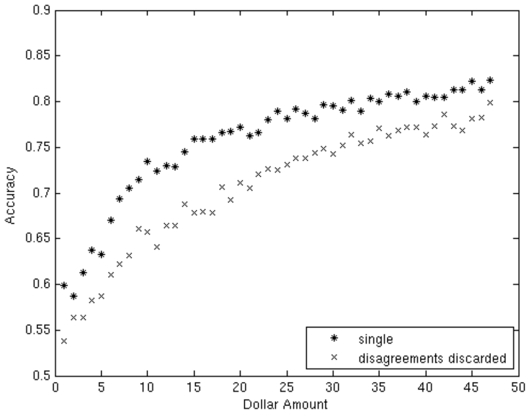

Figure 6: Performance of single annotated vs. double annotated data with disagreements discarded by amount invested for *to call*

Figures 7 and 8 present results averaged over all 215 verbs in the dataset. First, figure 7 shows the average accuracy over all verbs by amount invested. Again, these accuracy curves are not smooth because the verbs all have a different number of total instances. Hence, the accuracy values might jump or drop by a larger amount at the points where a given verb is no longer included in the average.

Toward the higher dollar amounts the curve is dominated by fewer and fewer verbs. As before, we only display the results for investments of up to $60.

The average difference in accuracy for Figure 7 across all amounts of investment is 2.32%.

Figure 8 presents the average accuracy relative to the percent of the total cost to single annotate all of the instances for a verb. The accuracy at a given percent of total investment was interpolated for each verb and then averaged over all of the verbs.

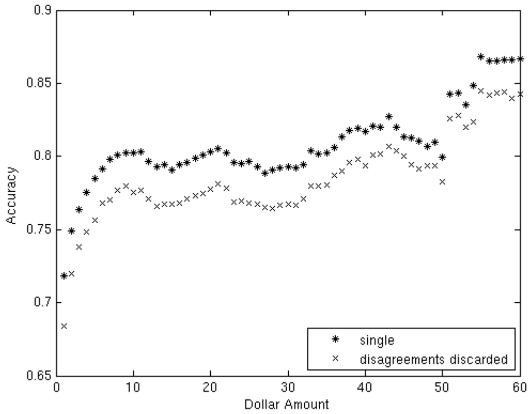

Figure 7: Average performance of single annotated vs. double annotated data with disagreements discarded by amount invested

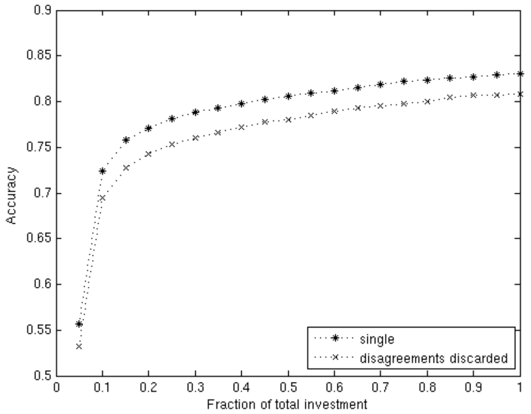

Figure 8: Average performance of single annotated vs. adjudicated data by fraction of total investment

The average difference in accuracy for Figure 8 across all amounts of investment is 2.51%.

Figures 9 and 10 present this information as a reduction in error rate for single annotation relative to full adjudication.

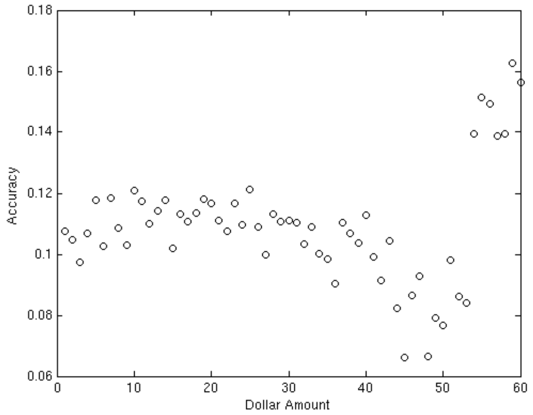

Figure 9: Reduction in error rate from adjudication to single annotation scenario based on results in Figure 7

The relative reduction in error rate averaged over all investment amounts in Figure 9 is 10.88%.

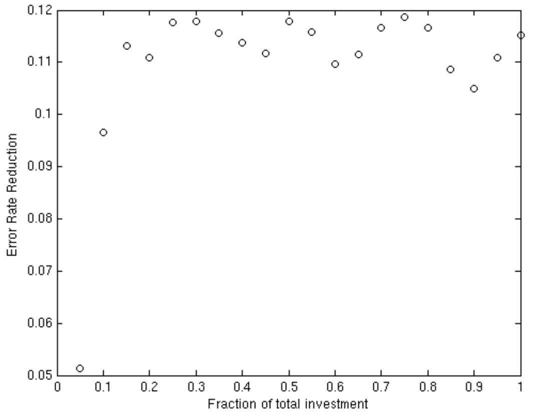

Figure 10: Reduction in error rate from adjudication to single annotation scenario based on results in Figure 8

The average relative reduction in error rate over the fractions of total investment in Figure 10 is 10.97%.

### 4.3 Discussion

At every amount of investment, our supervised WSD system performs better when trained on single annotated data comparing to double annotated data with discarded cases of disagreements.

The maximum annotation investment amount for each verb is the cost of single annotating all of its instances. When the system is trained on the amount of double annotated data possible at this investment, its accuracy is 80.78% (Figure 8). When trained on single annotated data, the system reaches the same accuracy much earlier, at approximately 52% of the total investment. When trained on the entire available single annotated data, the system attains an accuracy of 82.99%, an 11.5% relative reduction in error rate compared to the same system trained on the double annotated data obtained for the same cost.

The average accuracy of the single annotation scenario outperforms the double annotated with disagreements discarded scenario at every dollar amount investigated.

While this empirical investigation only looked at verb WSD, it was performed using 215 distinct verb type datasets. These verbs each have contextual features that are essentially unique to that verb type and consequently, 215 distinct classifiers, one per verb type, are trained. Hence, these could loosely be considered 215 distinct annotation and classification tasks.

The fact that for the 215 classification tasks the single annotation scenario on average performed better than the discard disagreements scenario of Beigman and Klebanov (2009) strongly suggests that, while it is theoretically possible for annotation bias to, in turn, bias a classifier's learning, it is more likely that you will achieve better results by training on the single annotated data.

It is still an open issue whether it is generally best to adjudicate disagreements in the test set or to throw them out as suggested by (Beigman Klebanov and Beigman, 2009).

## 5 Discussion and Future Work

We investigated 215 WSD classification tasks, comparing performance under three annotation scenarios each with the equivalent annotation cost, single annotation, double annotation with disagreements adjudicated, and double annotation with disagreements discarded. Averaging over the 215 classification tasks, the system trained on single annotated data achieved 10.0% and 11.5% relative reduction in error rates compared to training on the equivalent investment in adjudicated and disagreements discarded data, respectively. While we believe these results will generalize to other annotation tasks, this is still an open question to be determined by future work.

There are probably similar issues in what were considered fixed costs for the purposes of this paper. For example, it may be possible to train fewer annotators, and invest the savings into annotating more data. Perhaps more appropriately, it may be feasible to simply cut back on the amount of training provided per annotator and instead annotate more data.

On the other hand, when the unlabeled data is not freely obtainable, double annotation may be more suitable as a route to improving system performance. There may also be factors other than cost-effectiveness which make double annotation desirable. Many projects point to their ITA rates and corresponding kappa values as a measure of annotation quality, and of the reliability of the annotators (Artstein and Poesio, 2008). The OntoNotes project used ITA rates as a way of evaluating the clarity of the sense inventory that was being developed in parallel with the annotation. Lexical entries that resulted in low ITA rates were revised, usually improving the ITA rate. Calculating these rates requires double-blind annotation. Annotators who consistently produced ITA rates lower than average were also removed from the project. Therefore, caution is advised in determining when to dispense with double annotation in favor of more cost effective single annotation.

Double annotation can also be used to shed light on other research questions that, for example, require knowing which instances are ”hard.” That knowledge may help with designing additional, richer annotation layers or with cognitive science investigations into human representations of language.

Our results suggest that systems would likely benefit more from the larger training datasets that single annotation makes possible than from the less noisy datasets resulting from adjudication. Regardless of whether single or double annotation with adjudication is used, there will always be noise. Hence, we see the further investigation of algorithms that generalize despite the presence of noise to be critical to the future of computational linguistics. Humans are able to learn in the presence of noise, and our systems must follow suit.

## 6 Conclusion

Double annotated data contains less noise than single annotated data and thus improves the performance of supervised machine learning systems that are trained on a specific amount of data. However, double annotation is expensive and the alternative of single annotating more data instead is on the table for many annotation projects.

In this paper we compared the performance of a supervised machine learning system trained on double annotated data versus single annotated data obtainable for the same cost. Our results clearly demonstrate that single annotating more data can be a more cost-effective way to improve the system performance in the many cases where the unlabeled data is freely available and there are no other considerations that necessitate double annotation.

## 7 Acknowledgements

We gratefully acknowledge the support of the National Science Foundation Grant NSF-0715078, Consistent Criteria for Word Sense Disambiguation, and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, a subcontract from the BBN-AGILE Team. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

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

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *ACL-IJCNLP ’09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 280–287, Morristown, NJ, USA. Association for Computational Linguistics.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Comput. Linguist.*, 35(4):495–503.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what‘s in a name. *Mach. Learn.*, 34(1-3):211–231.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

J. Chen and M. Palmer. 2005. Towards robust high performance word sense disambiguation of english verbs using rich linguistic features. pages 933–944. Springer.

Jinying Chen, Dmitriy Dligach, and Martha Palmer. 2007. Towards large-scale high-performance english verb sense disambiguation by using linguistically motivated features. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 378–388, Washington, DC, USA. IEEE Computer Society.

Dmitriy Dligach and Martha Palmer. 2008. Novel semantic features for verb sense disambiguation. In *HLT '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 29–32, Morristown, NJ, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press Cambridge, MA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60, Morristown, NJ, USA. Association for Computational Linguistics.

A. Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.