```
    @inproceedings{Madnani:Boyd-Graber:Resnik-2010,
Title = {Measuring Transitivity Using Untrained Annotators},
Author = {Nitin Madnani and Jordan Boyd-Graber and Philip Resnik},
Booktitle = {Creating Speech and Language Data With Amazon's Mechanical Turk},
Year = {2010},
Location = {Los Angeles, CA},
Url = {http://cs.umd.edu/~jbg//docs/madnani-boyd-graber-turk-workshop.pdf},
}
```

Links:

- Data [http://cs.umd.edu/~jbg/downloads/mturk-transitivity.zip]

Measuring Transitivity Using Untrained Annotators

Nitin Madnani[a,b] **Jordan Boyd-Graber**[a] **Philip Resnik**[a,c]
[a]Institute for Advanced Computer Studies
[b]Department of Computer Science
[c]Department of Linguistics
University of Maryland, College Park
{nmadnani,jbg,resnik}@umiacs.umd.edu

Abstract

Hopper and Thompson (1980) defined a multi-axis theory of transitivity that goes beyond simple syntactic transitivity and captures how much "action" takes place in a sentence. Detecting these features requires a deep understanding of lexical semantics and real-world pragmatics. We propose two general approaches for creating a corpus of sentences labeled with respect to the Hopper-Thompson transitivity schema using Amazon Mechanical Turk. Both approaches assume no existing resources and incorporate all necessary annotation into a single system; this is done to allow for future generalization to other languages. The first task attempts to use language-neutral videos to elicit human-composed sentences with specified transitivity attributes. The second task uses an iterative process to first label the actors and objects in sentences and then annotate the sentences' transitivity. We examine the success of these techniques and perform a preliminary classification of the transitivity of held-out data.

Hopper and Thompson (1980) created a multi-axis theory of Transitivity[1] that describes the volition of the subject, the affectedness of the object, and the duration of the action. In short, this theory goes beyond the simple grammatical notion of transitivity (whether verbs take objects — transitive — or not — intransitive) and captures how much "action" takes place in a sentence. Such notions of Transitivity are not apparent from surface features alone; identical syntactic constructions can have vastly different Transitivity. This well-established linguistic theory, however, is not useful for real-world applications without a Transitivity-annotated corpus.

Given such a substantive corpus, conventional machine learning techniques could help determine the Transitivity of verbs within sentences. Transitivity has been found to play a role in what is called "syntactic framing," which expresses implicit sentiment (Greene and Resnik, 2009). In these contexts, the perspective or sentiment of the writer is reflected in the constructions used to express ideas. For example, a less Transitive construction might be used to deflect responsibility (e.g. "John was killed" vs. "Benjamin killed John").

In the rest of this paper, we review the Hopper-Thompson transitivity schema and propose two relatively language-neutral methods to collect Transitivity ratings. The first asks humans to generate sentences with desired Transitivity characteristics. The second asks humans to rate sentences on dimensions from the Hopper-Thompson schema. We then discuss the difficulties of collecting such linguistically deep data and analyze the available results. We then pilot an initial classifier on the Hopper-Thompson dimensions.

1 Transitivity

Table 1 shows the subset of the Hopper-Thompson dimensions of Transitivity used in this study. We excluded noun-specific aspects as we felt that these were well covered by existing natural language processing (NLP) approaches (e.g. whether the object / subject is person, abstract entity, or abstract concept is handled well by existing named entity recognition systems) and also excluded aspects which we felt had significant overlap with the dimensions we were investigating (e.g. affirmation and mode).

We also distinguished the original Hopper-Thompson "affectedness" aspect into separate "benefit" and "harm" components, as we suspect that these data will be useful to other applications such as sentiment analysis.

We believe that these dimensions of transitivity are simple and intuitive enough that they can be understood and labeled by the people on Amazon Mechanical Turk, a web service. Amazon Mechanical Turk (MTurk) allows individuals to post jobs on MTurk with a set fee that are then performed by workers on the Internet. MTurk connects workers to people with tasks and handles the coordination problems of payment and transferring data.

2 Experiments

Our goal is to create experiments for MTurk that will produce a large set of sentences with known values of Transitivity. With both experiments, we design the tasks to be as language independent as possible, thus not depending on language-specific preprocessing tools. This allows the data collection approach to be replicated in other languages.

[1] We use capital "T" to differentiate from conventional syntactic transitivity throughout the paper.

kinesis Sentences where movement happens are perceived to be more Transitive. “Sue jumped out of an airplane” vs. “The corporation jumped to a silly conclusion.”

punctuality Sentences where the action happens quickly are perceived to be more Transitive. “Caroline touched her ID card against the scanner to get through the locked door” vs. “I was touched by how much Irene helped me when I broke my leg.”

mode Sentences where there is no doubt about whether the action happened are perceived to be more Transitive. “Bob was too busy to fix the drain” vs. “Bob fixed the drain.”

affectedness Sentences where the object is more affected by the action are perceived to be more Transitive. “The St. Bernard saved the climber” vs. “Melanie looked at the model.”

volition Sentences where the actor chose to perform the action are perceived to be more Transitive. “Paul jumped out of the bushes and startled his poor sister” vs. “The picture startled George.”

aspect Sentences where the action is done to completion are perceived to be more Transitive. “Walter is eating the hamburger” vs. “Walter ate the pudding up.”

Table 1: The Hopper-Thompson dimensions of transitivity addressed in this paper. In experiments, “affectedness” was divided into “harm” and “benefit.”

2.1 Elicitation

The first task is not corpus specific, and requires no language-specific resources. We represent verbs using videos (Ma and Cook, 2009). This also provides a form of language independent sense disambiguation. We display videos illustrating verbs (Figure 1) and ask users on MTurk to identify the action and give nouns that can do the action and — in a separate task — the nouns that the action can be done to. For quality control, Turkers must match a previous Turker’s response for one of their answers (a la the game show “Family Feud”).

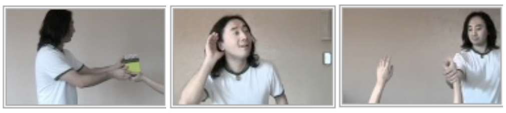
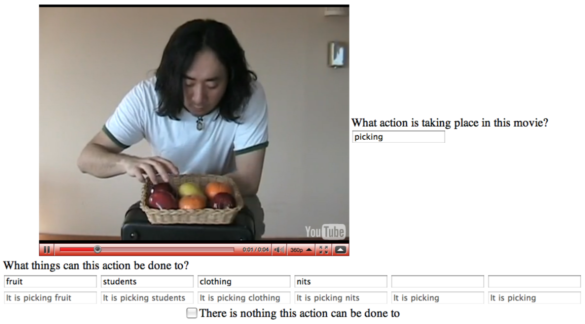


Figure 1: Stills from three videos depicting the verbs “receive,” “hear,” and “help.”

We initially found that subjects had difficulty distinguishing what things could do the action (subjects) vs. what things the action could be done to (objects). In order to suggest the appropriate syntactic frame, we use javascript to form their inputs into protosentences as they typed. For example, if they identified an action as “picking” and suggested “fruit” as a possible object, the protosentence “it is picking fruit” is displayed below their input (Figure 2). This helped ensure consistent answers. The subject and object tasks were done separately, and for the object task, users were allowed to say that there is nothing the action can be done to (for example, for an intransitive verb).

These subjects and objects we collected were then used as inputs for a second task. We showed workers videos with potential subjects and objects and asked them to create pairs of sentences with opposite Transitivity attributes. For example, *Write a sentence where the thing to which the action is done benefits* and *Write a sentence where the thing to which the action is done is not affected by the action.* For both sides of the Transitivity dimension, we allowed users to say that writing such a sentence is impossible. We discuss the initial results of this task in Section 3.



What action is taking place in this movie?
picking

What things can this action be done to?

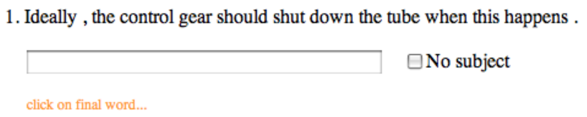
fruit	students	clothing	nits		
It is picking fruit	It is picking students	It is picking clothing	It is picking nits	It is picking	It is picking

□ There is nothing this action can be done to

Figure 2: A screenshot of a user completing a task to find objects of a particular verb, where the verb is represented by a film. After the user has written a verb and a noun, a protosentence is formed and shown to ensure that the user is using the words in the appropriate roles.

2.2 Annotation

Our second task—one of annotation—depends on having a corpus available in the language of interest. For concreteness and availability, we use Wikipedia, a free multilingual encyclopedia. We extract a large pool of sentences from Wikipedia containing verbs of interest. We apply light preprocessing to remove long, unclear (e.g. starting with a pronoun), or uniquely Wikipedian sentences (e.g. very short sentences of the form “See *List of Star Trek Characters*”). We construct tasks, each for a single verb, that ask users to identify the subject and object for the verb in randomly selected sentences.[2] Users were prompted by an interactive javascript guide (Figure 3) that instructed them to click on the first word of the subject (or object) and then to click on the last word that made up the subject (or object). After they clicked, a text box was automatically populated with their answer; this decreased errors and made the tasks easier to finish. For quality control, each HIT has a simple sentence where subject and object were already determined by the authors; the user must match the annotation on that sentence for credit. We ended up rejecting less than one percent of submitted hits.



1. Ideally , the control gear should shut down the tube when this happens .

□No subject

click on final word...

Figure 3: A screenshot of the subject identification task. The user has to click on the phrase that they believe is the subject.

Once objects and subjects have been identified, other users rate the sentence's Transitivity by answering the following questions like, where $VERB represents the verb of interest, $SUBJ is its subject and $OBJ is its object[3]:

- **Aspect**. After reading this sentence, do you know that $SUBJ is done $VERBing?
- **Affirmation**. From reading the sentence, how certain are you that $VERBing happened?
- **Benefit**. How much did $OBJ benefit?
- **Harm**. How much was $OBJ harmed?
- **Kinesis**. Did $SUBJ move?
- **Punctuality**. If you were to film $SUBJ's act of $VERBing in its entirety, how long would the movie be?
- **Volition**. Did the $SUBJ make a conscious choice to $VERB?

The answers were on a scale of 0 to 4 (higher numbers meant the sentence evinced more of the property in question), and each point in the scale had a description to anchor raters and to ensure consistent results.

2.3 Rewards

Table 2 summarizes the rewards for the tasks used in these experiments. Rewards were set at the minimal rate that could attract sufficient interest from users. For the "Video Elicitation" task, where users wrote sentences with specified Transitivity properties, we also offered bonuses for clever, clear sentences. However, this was our least popular task, and we struggled to attract users.

3 Results and Discussion

3.1 Creative but Unusable Elicitation Results

We initially thought that we would have difficulty coaxing users to provide full sentences. This turned out not to be the case. We had no difficulty getting (very imaginative) sentences, but the sentences were often inconsistent with the Transitivity aspects we are interested in. This shows both the difficulty of writing concise instructions for non-experts and the differences between everyday meanings of words and their meaning in linguistic contexts.

For example, the "volitional" elicitation task asked people to create sentences where the subject made a conscious decision to perform the action. In the cases where we asked users to create sentences where the subject did not make a conscious decision to perform an action, almost all of the sentences created by users focused on sentences where a person (rather than employ other tactics such as using a less individuated subject, e.g. replacing "Bob" with "freedom") was performing the action and was coerced into doing the action. For example:

- Sellers often give gifts to their clients when they are trying to make up for a wrongdoing.
- A man is forced to search for his money.
- The man, after protesting profusely, picked an exercise class to attend
- The vegetarian Sherpa had to eat the pepperoni pizza or he would surely have died.

While these data are likely still interesting for other purposes, their biased distribution is unlikely to be useful for helping identify whether an arbitrary sentence in a text expresses the volitional Transitivity attribute. The users prefer to have an animate agent that is compelled to take the action rather than create sentences where the action happens accidentally or is undertaken by an abstract or inanimate actor.

Similarly, for the aspect dimension, many users simply chose to represent actions that had not been completed using the future tense. For the kinesis task, users displayed amazing creativity in inventing situations where movement was correlated with the action. Unfortunately, as before, these data are not useful in generating predictive features for capturing the properties of Transitivity.

We hope to improve experiments and instructions to better align everyday intuitions with the linguistic properties of interest. While we have found that extensive directions tend to discourage users, perhaps there are ways incrementally building or modifying sentences that would allow us to elicit sentences with the desired Transitivity properties. This is discussed further in the conclusion, Section 4.

[2] Our goal of language independence and the unreliable correspondence between syntax and semantic roles precludes automatic labeling of the subjects and objects.

[3] These questions were developed using Greene and Resnik's (2009) surveys as a foundation.

Task	Questions / Hit	Pay	Repetition	Tasks	Total
Video Object	5	0.04	5	10	$2.00
Video Subject	5	0.04	5	10	$2.00
Corpus Object	10	0.03	5	50	$7.50
Corpus Subject	10	0.03	5	50	$7.50
Video Elicitation	5	0.10	2	70	$14.00
Corpus Annotation	7	0.03	3	400	$36.00
Total					$69.00

Table 2: The reward structure for the tasks presented in this paper (not including bonuses or MTurk overhead). "Video Subject" and "Video Object" are where users were presented with a video and supplied the subjects and objects of the depicted actions. "Corpus Subject" and "Corpus Object" are the tasks where users identified the subject and objects of sentences from Wikipedia. "Video Elicitation" refers to the task where users were asked to write sentences with specified Transitivity properties. "Corpus Annotation" is where users are presented with sentences with previously identified subjects and objects and must rate various dimensions of Transitivity.

3.2 Annotation Task

For the annotation task, we observed that users often had a hard time keeping their focus on the words in question and not incorporating additional knowledge. For example, for each of the following sentences:

- Bonosus dealt with the eastern cities so harshly that his **severity** was remembered centuries later .
- On the way there, however, Joe and Jake pick another **fight** .
- The Black Sea was a significant naval theatre of World War I and saw both **naval and land battles** during World War II .
- Bush claimed that Zubaydah gave **information** that lead to al Shibh 's capture .

some users said that the objects in **bold** were greatly harmed, suggesting that users felt even abstract concepts could be harmed in these sentences. A rigorous interpretation of the affectedness dimension would argue that these abstract concepts were incapable of being harmed. We suspect that the negative associations (severity, fight, battles, capture) present in this sentence are causing users to make connections to harm, thus creating these ratings.

Similarly, world knowledge flavored other questions, such as kinesis, where users were able to understand from context that the person doing the action probably moved at some point near the time of the event, even if movement wasn't a part of the act of, for example, "calling" or "loving."

3.3 Quantitative Results

For the annotation task, we were able to get consistent ratings of transitivity. Table 3 shows the proportion of sentences where two or more annotators agreed on the a Transitivity label of the sentences for that dimension. All of the dimensions were significantly better than random chance agreement (0.52); the best was harm, which has an accessible, clear, and intuitive definition, and the worst was kinesis, which was more ambiguous and prone to disagreement among raters.

Dimension	Sentences with Agreement
HARM	0.87
AFFIRMATION	0.86
VOLITION	0.86
PUNCTUALITY	0.81
BENEFIT	0.81
ASPECT	0.80
KINESIS	0.70

Table 3: For each of the dimensions of transitivity, the proportion of sentences where at least two of three raters agreed on the label. Random chance agreement is 0.52.

Figure 4 shows a distribution for each of the Transitivity data on the Wikipedia corpus. These data are consistent with what one would expect from random sentences from an encyclopedic dataset; most of the sentences encode truthful statements, most actions have been completed, most objects are not affected, most events are over a long time span, and there is a bimodal distribution over volition. One surprising result is that for kinesis there is a fairly flat distribution. One would expect a larger skew toward non-kinetic words. Qualitative analysis of the data suggest that raters used real-world knowledge to associate motion with the context of actions (even if motion is not a part of the action), and that raters were less confident about their answers, prompting more hedging and a flat distribution.

3.4 Predicting Transitivity

We also performed an set of initial experiments to investigate our ability to predict Transitivity values for held out data. We extracted three sets of features from the sentences: lexical features, syntactic features, and features derived from WordNet (Miller, 1990).

Lexical Features A feature was created for each word in a sentence after being stemmed using the Porter stem-

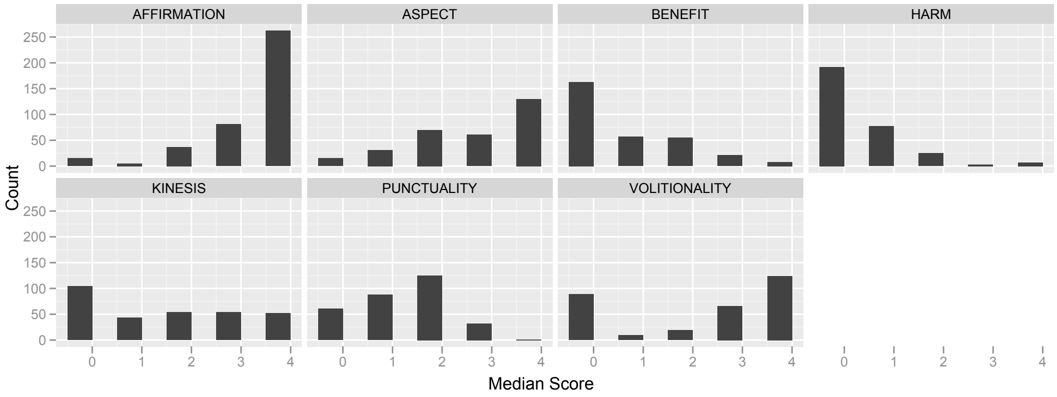


Figure 4: Histograms of median scores from raters by Transitivity dimension. Higher values represent greater levels of Transitivity.

mer (Porter, 1980).

Syntactic Features We parsed each sentence using the Stanford Parser (Klein and Manning, 2003) and used heuristics to identify cases where the main verb is transitive, where the subject is a nominalization (e.g. "running"), or whether the sentence is passive. If any of these constructions appear in the sentence, we generate a corresponding feature. These represent features identified by Greene and Resnik (2009).

WordNet Features For each word in the sentence, we extracted all the possible senses for each word. If any possible sense was a hyponym (i.e. an instance of) one of: *artifact*, *living thing*, *abstract entity*, *location*, or *food*, we added a feature corresponding to that top level synset. For example, the string "Lincoln" could be an instance of both a *location* (Lincoln, Nebraska) and a *living thing* (Abe Lincoln), so a feature was added for both the *location* and *living thing* senses. In addition to these noun-based features, features were added for each of the possible verb frames allowed by each of a word's possible senses (Fellbaum, 1998).

At first, we performed simple 5-way classification and found that we could not beat the most frequent class baseline for any dimension. We then decided to simplify the classification task to make binary predictions of low-vs-high instead of fine gradations along the particular dimension. To do this, we took all the rated sentences for each of the seven dimensions and divided the ratings into low (ratings of 0-1) and high (ratings of 2-4) values for that dimension. Table 4 shows the results for these binary classification experiments using different classifiers. All of the classification experiments were conducted using the Weka machine learning toolkit (Hall et al., 2009) and used 10-fold stratified cross validation.

Successfully rating Transitivity requires knowledge beyond individual tokens. For example, consider kinesis. Judging kinesis requires lexical semantics to realize whether a certain actor is capable of movement, pragmatics to determine if the described situation permits movement, and differentiating literal and figurative movement.

One source of real-world knowledge is WordNet; adding some initial features from WordNet appears to help aid some of these classifications. For example, classifiers trained on the volitionality data were not able to do better than the most frequent class baseline before the addition of WordNet-based features. This is a reasonable result, as WordNet features help the algorithm generalize which actors are capable of making decisions.

4 Conclusion

We began with the goal of capturing a subtle linguistic property for which annotated datasets were not available. We created a annotated dataset of 400 sentences taken from the real-word dataset Wikipedia annotated for seven different Transitivity properties. Users were able to give consistent answers, and we collected results in a manner that is relatively language independent. Once we expand and improve this data collection scheme for English, we hope to perform similar data collection in other languages. We have available the translated versions of the questions used in this study for Arabic and German.

Our elicitation task was not as successful as we had hoped. We learned that while we could form tasks using everyday language that we thought captured these subtle linguistic properties, we also had many unspoken assumptions that the creative workers on MTurk did not necessarily share. As we articulated these assumptions in increasingly long instruction sets to workers, the sheer size of the instructions began to intimidate and scare off workers.

While it seems unlikely we can strike a balance that will give us the answers we want with the elegant instructions that workers need to feel comfortable for the tasks as we currently defined them, we hope to modify the task to embed further linguistic assumptions. For example, we hope to pilot another version of the elicita-

Dimension	Makeup	Classifier Accuracy						
		Baseline	NB		VP		SVM	
			-WN	+WN	-WN	+WN	-WN	+WN
HARM	269/35	**88.5**	83.9	84.9	87.2	87.8	88.5	88.5
AFFIRMATION	380/20	**95.0**	92.5	92.0	94.3	95.0	95.0	95.0
VOLITION	209/98	68.1	66.4	69.4	67.1	**73.3**	68.1	68.1
PUNCTUALITY	158/149	51.5	59.6	**61.2**	57.0	59.6	51.5	51.5
BENEFIT	220/84	72.4	69.1	65.1	**73.4**	71.4	72.4	72.4
ASPECT	261/46	**85.0**	76.5	74.3	81.1	84.7	85.0	85.0
KINESIS	160/147	52.1	**61.2**	**61.2**	56.4	60.9	52.1	52.1

Table 4: The results of preliminary binary classification experiments for predicting various transitivity dimensions using different classifiers such as Naive Bayes (NB), Voted Perceptron (VP) and Support Vector Machines (SVM). Classifier accuracies for two sets of experiments are shown: without WordNet features (-WN) and with WordNet features (+WN). The baseline simply predicts the most frequent class. For each dimension, the split between low Transitivity (rated 0-1) and high Transitivity (rated 2-4) is shown under the "Makeup" column. All reported accuracies are using 10-fold stratified cross validation.

tion task where workers modify an existing sentence to change one Transitivity dimension. Instead of reading and understanding a plodding discussion of potentially irrelevant details, the user can simply see a list of sentence versions that are not allowed.

Our initial classification results suggest that we do not yet have enough data to always detect these Transitivity dimensions from unlabeled text or that our algorithms are using features that do not impart enough information. It is also possible that using another corpus might yield greater variation in Transitivity that would aid classification; Wikipedia by design attempts to keep a neutral tone and eschews the highly charged prose that would contain a great deal of Transitivity.

Another possibility is that, instead of just the Transitivity ratings alone, tweaks to the data collection process could also help guide classification algorithms (Zaidan et al., 2008). Thus, instead of clicking on a single annotation label in our current data collection process, Turkers would click on a data label *and* the word that most helped them make a decision.

Our attempts to predict Transitivity are not exhaustive, and there are a number of reasonable algorithms and resources which could also be applied to the problem; for example, one might expect semantic role labeling or sense disambiguation to possibly aid the prediction of Transitivity. Determining which techniques are effective and the reasons why they are effective would aid not just in predicting Transitivity, which we believe to be an interesting problem, but also in *understanding* Transitivity.

Using services like MTurk allows us to tighten the loop between data collection, data annotation, and machine learning and better understand difficult problems. We hope to refine the data collection process to provide more consistent results on useful sentences, build classifiers, and extract features that are able to discover the Transitivity of unlabeled text. We believe that our efforts will help cast an interesting aspect of theoretical linguistics into a more pragmatic setting and make it accessible for use in more practical problems like sentiment analysis.

References

C. Fellbaum, 1998. *WordNet : An Electronic Lexical Database*, chapter A semantic network of English verbs. MIT Press, Cambridge, MA.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Paul J. Hopper and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language*, (56):251–299.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the Association for Computational Linguistics*, pages 423–430.

Xiaojuan Ma and Perry R. Cook. 2009. How well do visual verbs work in daily communication for young and old adults? In *international conference on Human factors in computing systems*, pages 361–364.

George A. Miller. 1990. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2008. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NIPS*2008 Workshop on Cost Sensitive Learning*, Whistler, BC, December. 10 pages.