Bootstrapping Relation Extraction Using Parallel News Articles

Michael Glass
Department of Computer Science
University of Texas at Austin
mrglass@cs.utexas.edu

Ken Barker
Department of Computer Science
University of Texas at Austin
kbarker@cs.utexas.edu

Abstract

Relation extraction is the task of finding entities in text connected by semantic relations. Bootstrapping approaches to relation extraction have gained considerable attention in recent years. These approaches are built with an underlying assumption, that when a pair of words is known to be related in a specific way, sentences containing those words are likely to express that relationship. Therefore, sentences containing the pair of words may be used as training data for the relation extractor. We test this assumption for various relations drawn from two domains, using parallel and non-parallel corpora of news articles. We find that the assumption holds with substantially greater probability for parallel corpora.

1 Background and Related Work

Due to the enormous variety of expression in natural language, semantic structures can surface in many ways. Finding each of these ways manually, either by annotating or constructing interpretation rules by hand is impractical. One goal of DARPA's Machine Reading Project is to minimize this human effort when moving to new domains. Semi-supervised bootstrapping methods in relation extraction are a natural choice because of their ability to benefit from a small amount of manual effort and a large, unannotated corpus.

The Machine Reading Project's goals are much deeper than just relation extraction. Our approach is component-wise, and we consider relation extraction to be one component of the machine reading task. Other components include word-to-concept mapping including word sense disambiguation and named entity recognition, and co-reference resolution including event co-reference and cross-document co-reference. These components are not totally separable, each one constrains and informs the others. We use Markov Logic Networks [Domingos *et al.*, 2006] (MLNs) to perform joint inference over the outputs of the natural language processing components and the background knowledge.

Algorithm 1 gives a general outline of relation extraction bootstrapping. This algorithm bootstraps an extractor E for a single relation, although in some cases it may be beneficial to bootstrap multiple relations together [Carlson *et al.*, 2009]. It begins with a corpus, C, and a seed set of related pairs of instances, R, for example $\{(Cavaliers, 109), (Suns, 91)...\}$ for the *teamScore* relation. Alternatively, we may obtain seeds from a high precision (but low recall) relation extractor, in which case we first run the relation extractor over the corpus to obtain R. The first step of the bootstrapping loop is to label all occurrences of the pairs from R in C. Then an extractor is trained to identify the labeled sentences as examples of the target relation. The process repeats with the trained extractor providing an expanded set R.

Algorithm 1 A general relation bootstrapping algorithm

Input:
R: A set of instance pairs for a target relation.
C: An unlabeled corpus.
Output:
E: The trained relation extractor.
Procedure:

```
repeat
    O ← LabelOccurrences(R, C)
    E ← TrainExtractor(O)
    R_prev ← R
    R ← RunExtractor(E, C)
until R = R_prev
return E
```

In a seminal work, Hearst [1992] demonstrated a means of finding hyponyms from text using an initial set of simple syntactic patterns and a semi-automatically discovered set of additional syntactic patterns. Many researchers have also used parallel (or comparable) corpora to find paraphrases [Barzilay and Lee, 2003]. Shinyama et al. [2002] used parallel news stories to extract dependency paths with similar meanings based on shared named entity fillers.

A related idea, distant supervision [Mintz *et al.*, 2009], uses a database of known facts to label a separate corpus of natural language. Rather than beginning with a small number of seed instances, distant supervision uses a large database such as Freebase [Bollacker *et al.*, 2008] containing hundreds of thousands of entities. This approach operates not only on named entities but also nominals. For example, the */people/person/profession* relation relates a (named) person to a profession.

Bunescu and Mooney [2007] used a method of multiple instance learning to train a relation extractor using a similar idea. Though there was no bootstrapping, the system relied on the fact that a large enough bag of sentences containing two named entities known to be in relation will contain a sentence stating that relation. Their system began with sets of positive and negative pairs for each relation. Many sentences relating the positive pairs do not actually state the relation and none of the sentences relating the negative pairs state the relation. Therefore the penalty for misclassifying positive examples was set to be one ninth the penalty for misclassifying negative examples. This parameter was set by trial and error. They did not employ a parallel corpus.

Unlike previous work we consider not just entity-entity relations but also entity-event relations such as *teamInGame*. In these cases the event is almost never represented by a named entity and is often not represented by a noun phrase at all. Often it is a verb that refers to an event, such as "meet", "play" or "win" for a game event. While previous work used either a parallel corpus or a non-parallel corpus, we use both with a common set of relations.

In the context of pure relation extraction, extracting a instance pair for *teamInGame* such as $(Packers, played)$ has little value. In isolation all we can conclude is that the Green Bay Packers have played at least one game. However, in the context of Machine Reading, labeling two instances in a sentence as related by a specific relation can have much higher value. A sentence such as "The Packers played the Steelers" may yield the the two extractions $teamInGame(Packers, played)$ and $teamInGame(Steelers, played)$ where the two extractions are connected by the single instance mention of "played". Event co-reference may also enable the system to determine the date of the game referenced by "played" if it is stated in another sentence.

1.1 The Assumption of Bootstrapping

Algorithm 1 relies on the implicit assumption that when we encounter a sentence containing a pair of words known to be in a specific semantic relation, that sentence is likely to express that relation. This assumption has been at least partially validated by the success of the NLP bootstrapping methods that depend on it.

The focus of this paper is more narrow than bootstrapping relation extraction. We evaluate only a single function in the general bootstrapping algorithm: LABELOCCURRENCES. By determining its precision, we measure the degree to which the implicit bootstrapping assumption holds for various relations in parallel and non-parallel corpora.

2 Automatically Constructed Parallel Corpora

The Machine Reading Project motivated relation extraction on two domains: National Football League (NFL) and Intelligence Community (IC). The relations in the NFL ontology were general enough to apply to any team sport, so we will often refer to this as the Sports domain.

The abundance of news articles reporting on the same event makes parallel or comparable newswire corpora an attractive resource. Dolan et al. built an aligned corpus [2004], useful in paraphrase acquisition research, from news articles.

We built a large parallel corpus for both the Sports domain and the IC domain from clustered newswire. We used Google News[1] to locate and cluster articles describing the same story. Each document cluster covers a single news story, often a single event such as a game or terrorist attack. By searching Google News with domain relevant keywords, a set of results similar to Figure 1 is retrieved. By following the link to "all 285 articles" a set of documents all describing the same event can be gathered. Because both the relevance to the cluster and the quality (pagerank) of the articles decline with the number of search results retrieved, we limited the number of articles to the top ranked one third, or at most 100.


Washington Post

☆ **NBA** roundup: Monday's action on the boards
The Canadian Press - 5 hours ago
On a night the New York Knicks celebrated their 1970 **NBA** championship team, it was Andrew Bogut who stole the show. Bogut had 24 points and 20 rebounds in ...
On a Night to Savor Knicks Glory, the Present Disappoints Again New York Times
NBA New York Knicks' seventh losing streak The Daily Inquirer
Bogut, Bucks too much for Knicks on Legends Night NBA.com
UPI.com - Sportsbook Betting Promotions
all 285 news articles » Email this story

Figure 1: Example of a Google News article cluster

Each corpus is approximately half a gigabyte of text, gathered over about four months. On each day the corpus gathering software downloaded the clusters for the top ranked stories for each search term. The sports corpus was constructed by querying for news articles with the keywords: "NFL", "NBA", "NCAA football", and "MLB". The IC corpus was constructed from the keywords: "intelligence community", "conflict region", "Al Qaeda", "Taliban", "foreign election", and "insurgent". The sports corpus contains 145,000 documents across 3861 clusters, with an average of 37.5 documents per cluster. The IC corpus is similar, with 130,000 documents, 3114 clusters and 41.5 documents per cluster.

The news articles were automatically cleaned of the boilerplate typical of web articles and stripped of all HTML. The resulting text documents were segmented into sentences using Stanford's [Klein and Manning, 2003] sentence segmenter.

Because of the way the corpus was gathered, we know only that the documents inside each cluster were considered to be about the same story by Google News. We do not know that documents in other clusters are not about the same story. Often, multiple consecutive days will have articles about the same event. Given our architecture, these would appear in different clusters.

3 Experimental Setup

To determine the potential value of a parallel corpus for bootstrapping relation extractors, the experiment measures how consistently a relation holds between two seed instances within related news stories and across unrelated stories. Table

[1] http://news.google.com/

Relation	*Gloss*	*Search Pattern*
NFL (Sports) Relations		
gameDate	The game referred to as 'x' was played on y	" on "
gameWinner	The team, x, won the game referred to as 'y'	" won "
gameLoser	The team, x, lost the game referred to as 'y'	" lost "
teamInGame	The team, x, played in the game referred to as 'y'	" between "
teamScore	The team, x, scored y points	" scored "
Intelligence Community Relations		
eventLocationGPE	The event, x, took place in y	" in "
hasCitizenship	The person, x, is a citizen of y	" citizen of "
injuringHumanAgent	The group, x, was responsible for the attack, y	" claimed responsibility "
isLedBy	The organization, x, is led by y	" headed by "

Table 1: Relations for filler extraction

1 shows the five relations from each domain we considered. The gloss is an English explanation of the relation, where x and y are the two related entities. The search pattern is a snippet of text we used to locate sentences in the corpus containing entities in the relationship of interest.

Using the search patterns we first gathered sentences that potentially expressed the relation. From these we manually located sentences that actually expressed the target relation and located the words that were related. We refer to the pairs of words in these sentences as the seed instances. These pairs would be the initial set R in algorithm 1.

Because all sentences in which the seed instances are syntactically related will be labeled as occurrences of the target relation in the next bootstrapping step, it is essential that the instances be suitably specific. If the seed instances are overly general, such as $(group, he)$ for *isLedBy*, the LABELOCCURRENCES process will have low precision. If the seed instances are too specific, such as $(The\ surging\ Los\ Angeles\ Lakers, Monday's\ game)$ for *teamInGame*, very few sentences will be labeled as occurrences of the relation. To balance these considerations, we adopted the following seed instance selection strategy. First, no pronouns were accepted. For teams, either the name or city was acceptable. For people, a last name or descriptive general noun like "defendant" was accepted. For organizations referred to by name, only the most descriptive portion of the name was retained, for example "Harvard" for "Harvard University".

To simulate the LABELOCCURRENCES function in algorithm 1, a simple extractor randomly selected additional sentences in the corpus that contained a pair of seed instances. For each seed instance pair, it selected ten sentences to evaluate the precision of LABELOCCURRENCES within a single document cluster, called the "inside cluster" sentences. Since there might not be a single document cluster that had ten occurrences of a seed instance pair, inside cluster sentences were added when there were at least two within a single cluster: the extractor continued randomly adding all matching sentences from suitable clusters until the number of sentences minus the number of document clusters they were drawn from equaled or exceeded ten.

This method is shown formally in algorithm 2. The function SENTENCESRELATING(si, c) finds all sentences in the

Algorithm 2 Gather inside cluster sentences

Input:
si: An instance pair.
C: The parallel corpus, composed of sets c of document clusters.
Output:
ICS_{si}: The inside cluster sentences for the instance pair si. It is a set of sets of sentences relating the two instances. Each contained set is drawn from the same document cluster.
Procedure:

```
ICS_si ← ∅
size ← 0
for c ∈ C do
  S ← SENTENCESRELATING(si, c)
  if |S| ≥ 2 then
    ICS_si ← ICS_si ∪ {S}
    size ← size + |S| − 1
  end if
  if size ≥ 10 then
    return ICS_si
  end if
end for
```

Algorithm 3 Gather outside cluster sentences

Input:
si: An instance pair.
C: The parallel corpus, composed of sets c of document clusters.
Output:
OCS_{si}: The outside cluster sentences for the instance pair si. It is a set of sentences relating the two instances.
Procedure:

```
OCS_si ← ∅
for c ∈ C do
  S ← SENTENCESRELATING(si, c)
  if |S| ≥ 1 then
    OCS_si ← OCS_si ∪ {RANDOMELEMENT(S)}
  end if
  if |OCS_si| ≥ 10 then
    return OCS_si
  end if
end for
```

document cluster c that relate the seed instance pair si. Note that although the for loop appears to be iterating over the corpus in a fixed order, the document clusters were sampled randomly without replacement. The sentence gathering algorithm is run for each seed instance pair from each target relation.

We also selected another ten sentences spread across document clusters to test the precision of LABELOCCURRENCES without document clustering, called the "outside cluster" sentences. Though somewhat obvious, for completeness, this is shown in algorithm 3.

We required the sentences found in this way to have a close syntactic relationship between the two seed instances. We measured this by parsing the sentences and rejecting any sentence where the words were separated by a dependency path of length greater than five.

4 Evaluation

To evaluate the precision of the LABELOCCURRENCES function, with and without document clustering, we used workers on Amazon's Mechanical Turk[2] service. Other researchers [Snow *et al.*, 2008] have shown that for some simple tasks in natural language processing, including textual entailment, Mechanical Turk non-experts can provide annotations with the same level of quality as expert-built gold standards.

For each relation a gloss was written with its meaning expressed in plain English (see Table 1). The found sentences, along with the filled gloss for the relation's meaning were the Human Intelligence Tests (HITs). Each HIT (the labeled occurrence) was given to three workers, to test agreement.

Figure 2 shows the instructions given to the workers. The workers were asked to mark the sentence-summary pair as correct or incorrect (and if incorrect, select a reason why). We provided example sentence-summary pairs to illustrate the three different types of incorrect summaries.

5 Results

The Fleiss' Kappa for Mechanical Turk worker agreement on correctness across all relations was 0.333 indicating "Fair agreement" [Landis and Koch, 1977]. The *gameWinner* and *gameLoser* relations had low precision ($< 20\%$) and low agreement (< 0.1) for both the inside and outside cluster sentences. We exclude these relations from further analysis.

Mechanical Turk workers were unanimous in about half of the HITs. For the others, we reviewed a sample of 25 HITs. Of those, we agreed with the majority of workers 19 times. In cases where workers disagreed, they often had legitimate reasons for their disagreements. Consider the sentence below along with the gloss of the hypothesized relation.

> Sistan-Baluchistan province a major transit route for narcotics has been hit by a string of attacks and kidnappings, that authorities blame on Jundollah.

The group, **Jundollah**, was responsible for the attack, **attack**

Answers	correct 1	incorrect 2

The workers were instructed to mark the partial summary

- Determine if the sentence states or strongly implies the partial summary.
- If the sentence does not actually state (or strongly imply) the relationship, it is incorrect, even if you know the partial summary is true.
- If it is incorrect, consider why it is not correct.
  - Is a word in bold not of the type stated?
    - Sentence: John **won** an award in **New York**.
    - Summary: The team, **New York**, won the game referred to as **won**.
    - Explanation: "New York" does not refer to a team and "won" does not refer to a game.
  - Is there no direct relationship between the words?
    - Sentence: Kobe Bryant of the Los Angeles **Lakers** will **play** in the All-Star Game.
    - Summary: The team, **Lakers**, played in the game referred to as **play**.
    - Explanation: There is only an indirect relationship between the Lakers and the playing event.
  - Is the relationship between the words different from the summary?
    - Sentence: President **Obama** spoke at **Harvard** today.
    - Summary: The person, **Obama**, attended the school, **Harvard**.
    - Explanation: The sentence says that Obama spoke at Harvard, not attended
- Tense (past, present, future) is unimportant.

Figure 2: The instructions for Mechanical Turk workers.

correct if it was stated or "strongly implied" by the sentence. Whether the blame of authorities is enough to strongly imply responsibility is obviously a subjective matter.

In order to conduct an error analysis, we asked the Mechanical Turk workers to provide one of three reasons in cases where the words were not in the relationship stated. Some errors may be much more problematic for bootstrapping than others. For example, if the relationship does not exist because the words are not of the types stated, then a sufficiently advanced word-to-concept component could filter the error. Unfortunately, the Fleiss' Kappa for worker agreement when including the reasons for incorrectness was only 0.18, indicating "Slight agreement" [Landis and Koch, 1977]. A quick sampling of the worker's reasons for incorrectness also showed low agreement with the author's judgements. Therefore, we leave error analysis to future work.

5.1 Effect of Parallel Corpus

Algorithm 4 shows the method of evaluating the inside cluster precision. To measure the inside cluster precision for a relation, it loops over all seed instance pairs for that relation and locates all sentence-summary pairs where the workers unanimously agreed the partial summary was correct. Then it tallies the majority opinion - whether correct or incorrect - on every other sentence-summary pair from the same cluster.

To measure outside cluster precision we simply gathered the majority opinion for every outside cluster sentence. For completeness we also show this very straightforward

[2] http://mturk.com

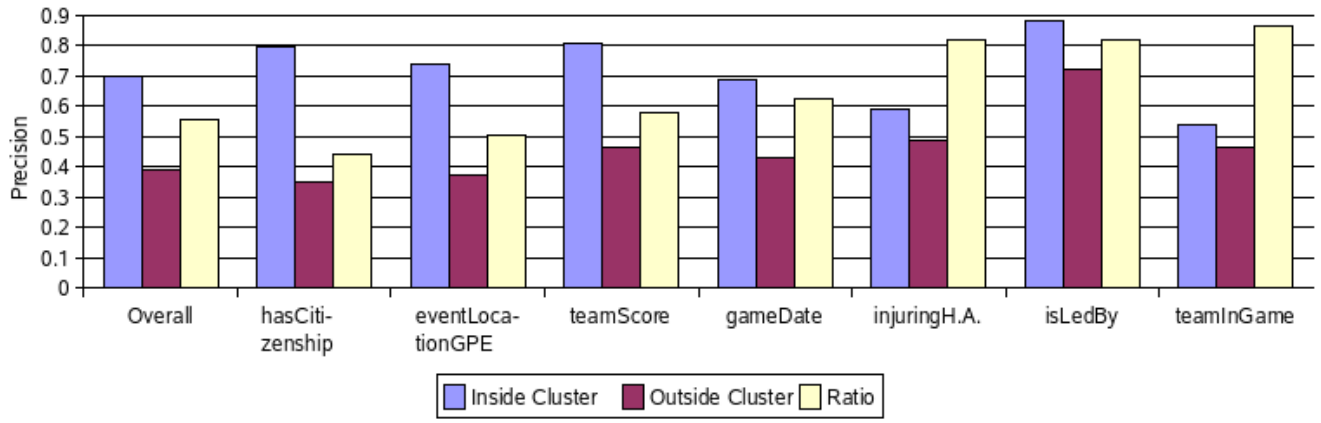

Figure 3: Impact of a parallel corpus on instance relationship precision

Algorithm 4 Scoring inside cluster sentences

Input:
R: A set of seed instance pairs for a target relation.
ICS_{si}: A set of sets of sentences relating the instance pair. Each contained set is drawn from the same document cluster.
Procedure:

```
for si ∈ R do
  for ics ∈ ICS_si do
    if ∃s_u ∈ ics : UNANIMOUSCORRECT(s_u) then
      for s_m ∈ ics/s_u do
        if MAJORITYCORRECT(s_m) then
          correct ← correct + 1
        else
          incorrect ← incorrect + 1
        end if
      end for
    end if
  end for
end for
```

Algorithm 5 Scoring outside cluster sentences

Input:
R: A set of seed instance pairs for a target relation.
OCS_{si}: A set of sentences relating the instance pair. Each sentence is drawn from a different document cluster.
Procedure:

```
for si ∈ R do
  for s ∈ OCS_si do
    if MAJORITYCORRECT(s) then
      correct ← correct + 1
    else
      incorrect ← incorrect + 1
    end if
  end for
end for
```

method in algorithm 5. Precision is then $correct/(correct + incorrect)$. The ratio reported in Figure 3 is the outside cluster precision divided by the inside cluster precision.

A one-tailed t-test confirms the difference between the inside and outside cluster precision is statistically significant, even adjusting for multiple testing ($p < 0.05$), for all relations but the last three: *injuringHumanAgent*, *isLedBy*, and *teamInGame*.

6 Conclusions

Figure 3 shows the importance of using a parallel corpus for at least some relations. Overall the inside cluster precision was 0.70 and the outside cluster precision was 0.39, suggesting a parallel corpus may lead to an error rate reduction of 50% in LABELOCCURRENCES. The effect varied considerably based on relation.

We expected a minimal impact for the *gameDate*, *teamInGame* and *injuringHumanAgent* relations and a substantial impact for the others. We reasoned that when the typical instance pairs for these relations are present together in a sentence it is rare that they are related in anyway except the target relation. Although we were unable to test some relations effectively, with the exceptions of *gameDate* and *isLedBy* our hypothesis was validated.

The dramatic drop from inside cluster to outside cluster precision for the *hasCitizenship* relation can be explained by the fact that, not coincidentally, a citizen of a nation is typically related to that nation in many ways. For *teamScore*, we see that within a cluster (usually about a single game) if the relationship holds between a team and a score in one sentence in that cluster, it is very likely to hold between all other syntactically related occurrences of those same instances. However, outside the cluster it is less than 50% likely that the relation holds. Often the score will actually be the *other* team's score.

For many cases the assumption of a single relation between seed instance pairs does not hold generally, but is reliable within a cluster of documents on a single story. Bootstrapping approaches to relation extraction can take advantage of this result by using parallel corpora for relations that can benefit.

Acknowledgments

The authors gratefully acknowledge the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0172. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

[Barzilay and Lee, 2003] Regina Barzilay and Lillian Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 16–23. Association for Computational Linguistics, 2003.

[Bollacker *et al.*, 2008] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[Bunescu and Mooney, 2007] Razvan Bunescu and Raymond Mooney. Learning to Extract Relations from the Web using Minimal Supervision. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 576–583, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[Carlson *et al.*, 2009] Andrew Carlson, Justin Betteridge, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupling semi-supervised learning of categories and relations. In *Proc. of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 1–9. Association for Computational Linguistics, 2009.

[Dolan *et al.*, 2004] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04: Proc. of the 20th International Conference on Computational Linguistics*, page 350. Association for Computational Linguistics, 2004.

[Domingos *et al.*, 2006] Pedro Domingos, Stanley Kok, Hoifung Poon, Matthew Richardson, and Parag Singla. Unifying logical and statistical ai. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 2–7. AAAI Press, 2006.

[Hearst, 1992] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc, of the 14th Conference on Computational Linguistics*, pages 539–545. Association for Computational Linguistics, 1992.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *ACL '03: Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics, 2003.

[Landis and Koch, 1977] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proc. 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 2*, ACL '09, pages 1003–1011. Association for Computational Linguistics, 2009.

[Shinyama *et al.*, 2002] Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. Automatic paraphrase acquisition from news articles. In *Proc. of the Second International Conference on Human Language Technology Research*, pages 313–318. Morgan Kaufmann Publishers Inc., 2002.

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