Annotating Legitimate Disagreement in Corpus Construction

Billy T. M. Wong
Department of Translation
The Chinese University of Hong Kong
Hong Kong
billy@arts.cuhk.edu.hk

Sophia Y. M. Lee
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hong Kong
ym.lee@polyu.edu.hk

Abstract

This paper addresses the resolution of inter-annotator disagreement in corpus construction. Given the consistency requirement which is regarded as a critical criterion of annotation quality, inter-annotator disagreement is usually considered harmful to the accuracy and reliability of annotation, and thus has to be resolved through various means. We claim that strictly adhering to consistency would also neglect the legitimate disagreement originating from ambiguity in natural languages. We highlight the values of preserving legitimate disagreement in annotation, and show that the possible problems resulting from inconsistency are avoidable. A preliminary annotation scheme is suggested for supporting multiple versions of annotation, without giving up the virtue of consistency.

1 Introduction

Annotation is an important stage in corpus development. It enriches a corpus by providing explicit representations of linguistic information encoded in the texts, which supports the empirical study of linguistic phenomena and the development of natural language processing techniques. Depending on the purpose of corpus construction, types of annotation may include syllable boundary, part-of-speech, lemma, syntactic structure, semantic field, anaphoric relation, and many others. The annotation process can be carried out manually by linguists or trained people, automatically by computer programs, or semi-automatically through automatic annotation plus human post-editing.

The quality of annotation must be maintained for reliable corpus analysis. This involves the criteria of accuracy and consistency. The former refers to the correctness of annotation in accordance with the specifications usually provided in the form of guidelines. The latter relates to the extent of which annotators agree in their judgments with themselves and each other. The accuracy and consistency of annotation are also believed to have a close relationship. If the judgments from two or more annotators are all correct, then in most cases they should also be consistent. Although this may not be true the other way round, it is a rare case that consistent judgments from multiple annotators are incorrect when the sample size is large enough. The assumption of a strong correlation between accuracy and consistency allows us to rely on either of these criteria for assessing the annotation quality. In practice, consistency, which is measured in terms of inter-annotator agreement coefficients such as Cohen's Kappa (Cohen, 1960), is more commonly used. The primary advantage of this attribute is cost-effectiveness in checking the correctness of annotations without any human effort and establishing a golden standard in annotation.

Thus, maintaining a strong inter-annotator agreement has become a high priority in managing an annotation project. It involves resolving disagreements through various means, which may include adjustment or deletion of discordant annotations. What has to be revised may even include the kinds of linguistic phenomena to annotate and the way they are annotated, in order to reduce the occurrence of inconsistent judgments.

We claim that such a practice, however, does not fully embrace the intent of corpus annotation. In particular, it neglects the fact that disagreement may be caused by ambiguity in natural languages, such that annotators can have different yet legitimate judgments on the same linguistic phenomenon. These judgments would incur the risk of missing out on other possible interpretations. Without disregarding the importance of consisten-

International Joint Conference on Natural Language Processing, pages 51–57,
Nagoya, Japan, 14-18 October 2013.

cy, we suggest ways to preserve such legitimate disagreements in corpus annotation.

2 Current Approaches of Resolving Disagreement

This section reviews current approaches of resolving inter-annotator disagreement in corpus annotation. It is worth nothing that none of them are typically used in isolation, but in conjunction with the others in the iterative process of disagreement resolution.

2.1 Annotation guideline

Annotation guidelines specify the detailed procedure to record the linguistic phenomena in question, serving as the standard for annotators to follow. It is regarded as the most important means of ensuring the annotation accuracy and consistency. Inter-annotator disagreement can be minimized by tightening up the guidelines, clearly restricting how every problematic case is handled, with positive and negative examples provided as references or used as the "default" option (Xia et al., 2000) to prevent annotators from making individual choices. In other words, despite the cases that the guidelines are misinterpreted or ignored by annotators, the occurrence of disagreement indicates a problem with the guidelines. Poesio and Artstein (2005) criticize such a view— that the problem would disappear when finding the "right" annotation scheme or concentrating on the "right" linguistic judgments— as being misguided. Such a practice has made inter-annotator agreement "an artifact of annotation scheme and procedure" (Alm, 2010). Zaenen (2006) notes that "it suffices that all annotators do the same thing. But even with full annotator agreement it is not sure that the task captures what was originally intended".

As a matter of fact, there are still cases where inter-annotator agreement remains mild even after extensive guideline revision and annotator training (Morgan et al., 2013). It is also argued that following a tight annotation scheme may lead to many marginal cases (i.e. false negatives (Morgan et al., 2013)) being unannotated. Furthermore, for annotations of linguistic phenomena which are fuzzy and ambiguous in nature such as language errors of non-native learners (Rosen et al., 2013), it is questionable whether all grey areas can be fully clarified. Sometimes an expression can be classified as one of the two or more categories. Although annotators can be instructed to persist in a certain choice given in the guideline for consistency purposes, it conceals the fact that an expression can be perceived differently by different language users, as commented in Rosen et al. (2013).

2.2 Expert adjudication

In case of disagreement, the final decision can be made by an expert who may be one of the annotators. S/he may have expertise in the subject matter, or be an experienced annotator.

The reliability of this approach is then completely reliant on the quality of the experts. For annotation of linguistic phenomena which are subjective in nature, it is argued that there is no real expert (Carletta, 1996), where no one interpretation can be deemed superior to the others. Hong and Baker (2011) also observe that sometimes the majority of annotators are simply right, while the experts are wrong.

2.3 Discussion

Once there is disagreement, it is common for annotators to compare their differences and attempt to arrive at the proper choice. Examples of such practices include the annotation of Chinese collocations (Xu et al., 2007), discourse anaphora (Dipper and Zinsmeister, 2009), prosodic breaks (Jung and Kwon, 2011), and appraisal expressions (Read and Carroll, 2012). Sometimes, the discussion simply reveals a misunderstanding of annotators or unclear instructions in the guidelines. Through discussion, it is also intended to arrive at a set of gold-standard annotation used for checking the accuracy of other annotators (Xue et al., 2002; Ruppenhofer et al., 2012).

2.4 Removal

Highly-ambiguous or marginal entries may be simply removed from the annotation. This approach is applied in Chen et al. (2009) and Lee et al. (2010) for identification and classification of Chinese emotion. In their work, what is regarded as an emotion entity is largely determined by keywords carrying different degrees of emotional intensity, with a set of keywords classified as carrying strong emotion and another classified as carrying weak emotion. A threshold is determined that only the keywords with emotional intensity above the threshold are included in the annotated set while the remaining are discarded.

2.5 Relaxed criteria

In contrast with the practice of having a tight annotation scheme, the strictness of criteria can also be relaxed so as to allow slightly different judgments to be regarded as the same. For instance, in Penn Chinese Treebank (Xue et al., 2002) the internal structure of the noun phrase (which is sometimes difficult to determine) is not annotated, in order to simplify the task without loss of information.

In the annotation of discourse relation (Miltsakaki et al., 2004) and opinion and emotion expression (Wiebe et al., 2005), the boundaries of relevant expression (e.g. phrase, higher verb, dependent clause, parenthetical, sentence) are hardly definitive. Annotators usually identify "partial overlaps", with common text span between the different selections. The kind of intersecting expressions can be regarded as agreeing tokens if the criteria are relaxed.

For labeling of linguistic phenomena such as word senses which constitute a hierarchical structure in themselves, it is not uncommon to have disagreement when the labels are assigned at the finest level. For this kind of annotation, inter-annotator agreement is reported (Webber et al., 2003; Duffield et al., 2007; Read and Carroll, 2012) to increase when relaxing the strictness of annotation— opting for an upper level label in case of multiple possible judgments at a concrete level.

2.6 Crowd wisdom

The prevalence of utilizing collective effort (e.g. Games with a Purpose, Amazon Mechanical Turk, or Wisdom of Crowds) for annotation in recent years has also brought forth the problem of consistency. Compared with the traditional approach which involves at most two to three well-trained annotators, the number of annotators who are usually non-expert can be much larger in the collaborative approach. Although it is shown in Snow et al. (2008) that annotated data obtained from non-experts is as good as those from trained experts, Dandapat et al. (2009) find that annotation quality also depends on the nature of task.

A number of strategies are suggested in Wang et al. (2013) to ensure annotation accuracy and consistency, including the use of acceptance rating threshold for annotator screening, agreement threshold for monitoring annotators' judgments, gold-standard questions to detect spam workers, and the reliance of other workers to rate the quality of initial worker annotation.

When there are a sufficient number of annotators, Hong and Baker (2011) find that simply relying on the majority may be enough for resolving disagreement. A case of more or less equal number of votes indicates real ambiguity in the provided options.

3 Ambiguity Revisited

As reviewed, nearly all current approaches of resolving disagreement are intended to arrive at a single final judgment for maintaining consistency. It is also noticed that disagreement is nearly inevitable when there is more than one annotator. As studied in Dandapat et al. (2009) and Cui and Chi (2013), there are four major causes of disagreement. Aside from human errors, vague guidelines and ignorance about the guidelines, disagreement can also be caused by the inherent ambiguity in languages where various interpretations are all plausible and legitimate. Such interpretive ambiguity is widely reported in various annotation projects involving different kinds of linguistic phenomenon, such as predicate-argument and coreference relations (Versley, 2006; Iida et al., 2007), prosodic breaks (Jung and Kwon, 2011), semantic roles (Ruppenhofer et al., 2012), language learner errors (Rosen et al., 2013), and many others.

As a natural characteristic in human languages, ambiguity is classified by Poesio and Artstein (2005) into explicit and implicit types. The former can be immediately perceived by annotators while the latter can only be revealed by comparing their annotations to find out the difference in their interpretations.

3.1 Explicit ambiguity

Explicit ambiguity is well-studied in various linguistic disciplines. Typically, many words in English can function as more than one part-of-speech. In the British National Corpus (BNC) a set of portmanteau tags is used for annotating such ambiguity. For example, the tagging "liked_VVD-VVN" means that the word "liked" can either be the past tense or past participle of a lexical verb. At the syntactic level, another example from BNC is provided in Leech and Eyes (1997) as:

The main global-warning gas [...] is carbon dioxide, given off by burning fossil fuels.

The last three words can serve either as a gerundival *-ing* clause ([Tg burning_VVG [N fossil_NN1 fuels_NN2 N]Tg]) or a noun phrase ([N burning_JJ [fossil_NN1 fuels_NN2]N]). Even though there are multiple analyses, human readers can usually infer the more appropriate one based on the context.

3.2 Implicit ambiguity

Implicit ambiguity poses more of a challenge to resolve in annotation. It leads to different interpretations, which are all plausible. An agreement between annotators may not be able to arrive at even after discussion.

The difficulty of annotating discourse features is a typical case of implicit ambiguity. Features such as politeness are context-dependent in nature where their identification causes more dispute than that of other linguistic phenomena. In the annotation of appraisal expressions, Read and Carroll (2012) notice that even though annotators are highly familiar with the appraisal theory, disagreement still occurs in their judgments, mostly in the acceptability of marginal cases. Some annotators only accept clear prototypical expressions while some are more tolerant of fuzzyness. Cui and Chi (2013) provide an example of annotating model expression in the Penn Chinese Treebank (Xia et al., 2000):

歐盟表示要進一步促進雙方在各領域的交流。

The word 要 (yao) can be used as a modal or an attitude verb (non-modal). Therefore in this example there are two possible interpretations:

(i) EU says that the two parties need to further promote their communication in various areas. (model)

(ii) EU says that (it) is willing to further promote the communication between the two parties in carious areas. (non-model)

Some kinds of annotation, such as word sense assignment, rely entirely on annotators' perception. Erk et al. (2009) explain the disagreement in word sense assignment through the perspective of human cognition. The categories in human mind are related to various strengths of closeness rather than clearcut boundaries. Some items are perceived as more typical than the others while some are borderline cases which are the source of disagreement. Thus in their practice of word sense judgment annotators are instructed to give graded ratings instead of binary choices. Quan and Ren (2009) also allow annotators to use their own intuition in identifying Chinese emotional words. Disagreement is found in the set of emotional words identified between two annotators (i and ii) in the following example.

今晨，當我沐浴著陽光前往會場時，腦中突然浮現出多年不用的優美詞語，那就是：秋高氣爽、金光璀璨。

(This morning, as I was walking to the venue, bathed in sunlight, some wonderful words that have not been used for many years crossed my mind, which are "the autumn sky is clear, the air is crisp" and "shinning with gold color".)

Emotional words identified (inconsistent choices are underlined):

(i) 陽光, 優美, 秋高氣爽, 金光(gold color), 璀璨

(ii) 沐浴(bath), 陽光, 優美, 秋高氣爽, 璀璨

The annotation of understudied linguistic phenomena suffers further from the lack of a well-developed supporting theory. Alm (2010) describe the annotation of the Affect expression. Given that Affect is still an understudied phenomenon in linguistics, there is a lack of consensus on how it can be modelled. Similarly, Jung and Kwon (2011) find the identification of prosodic breaks as a task without clear definition, but largely dependent on annotators' own perception and interpretation. In Morgan et al. (2013) it is found that in the annotation of social acts, the identification of their occurrence and boundaries is difficult. Annotators are only able to consistently agree to prototypical cases. Moreover, the labels of social acts they use for annotation do not have well-established prior definitions. Indeed, one of the goals of their annotation project is to develop a typology of social acts.

As in many annotation projects whose aim is to collect instances of a linguistic phenomenon for further study, the linguistic phenomenon in question may not have a well-established definition. In this case disagreement is inevitable. Every instance of this kind of disagreement represents one controversial yet plausible reading based on the limited understanding and imperfect theory of that linguistic phenomenon. Therefore, missing any potential instance, even marginal, is a loss because those controversial cases indicate the difficult part that current theory does not solve satisfactorily.

In such cases, it is less clear how a strong inter-annotator agreement which can be produced artificially contributes to the study of linguistic phenomenon in question. In contrast, there have been suggestions to collect ambiguous expressions for further studies. For example, Wiebe et al. (2005) categorizes instances of annotated data into two types: reliable/unreliable and easy/hard, under the assumption that easy items can be reliably annotated. The annotation of hard cases is unreliable due to inconsistency, but more valuable for theory development, as they indicate where current theory is having difficulty. Once the theory is improved to support resolution of those ambiguous hard cases, they can be included into the annotated dataset without going through the whole corpus again for their identification. Similarity, Versley (2006) contends that the labeling of ambiguities help raising annotators' awareness on them. Alm (2010) claims to resort to flexible acceptability to capture subjective language phenomena when the ground truth is not available yet. Stede and Huang (2012) also raise that instead of having the same phenomenon annotated many times, it is more important to focus on the interesting and more difficult phenomena in order to derive insights from them.

4 Preserving Legitimate Disagreement

Following the above discussion, this section discusses how legitimate disagreement can be preserved. We define legitimate disagreement as the difference in judgments caused by ambiguity in languages which cannot be clearly resolved by current linguistic theory. This reserves the possibility of finding a satisfactory resolution in future. It should also be clarified that preserving disagreement does not necessarily imply the abandonment of consistency. Consistency remains an indispensable criterion in corpus annotation. It is one of the key prerequisites for extracting linguistic knowledge, and for providing reliable data for training and testing of natural language processing technology.

The first step of preserving legitimate disagreement is to identify it. This involves its differentiation from other kinds of inconsistent judgments caused by human errors or vague guidelines. In general this step does not impose much extra burden on annotators, as resolution of inconsistency is already a regular task in corpus annotation. Furthermore, it is useful to have an annotation scheme for recording inconsistent judgments once classified as legitimate, rather than revising or deleting them.

A workable approach is to add an extra attribute to the annotation scheme to indicate the ambiguous status of an expression. Take the annotation of Chinese emotion expression as an example. It is a typical understudied language phenomenon without a well-developed theory and is highly dependent on human perception. The difficulties are first to identify words carrying emotional sense; and second, to categorize the emotion words into their corresponding emotion classes. In Chen et al. (2009) and Lee et al. (2010), five primary emotion classes are first predefined, including *happiness*, *sadness*, *fear*, *anger*, and *surprise*, and a set of emotion words identified. However, difficulty resides in assigning the exact ambiguous emotion words, such as 如意 (as one wishes), 害羞 (to be shy) and 為難 (to feel embarrassed/awkward) to an emotion class. More likely, each of these emotion words tends to belong to more than one emotion class in different contexts. Instead of simply removing these ambiguous emotion words from the annotation for the sake of maintaining consistency, we can use an attribute <confidence> together with a level scale to signal the confidence of the classes to which this emotion word belongs. Using a five-point scale [0,1,2,3,4] where level-0 refers to the most confident level and level-4 the least, an example of annotation can be:

嘉莉沒有參加他們的婚禮，他們對此很 <emotionword class='anger' confidence=1; class='sadness' confidence=4> 不高興 </emotionword> 。
(They were <emotionword class='anger' confidence=1; class='sadness' confidence=4> unhappy</emotionword> when Carrie did not come to their wedding.)

In this example the expression 不高興 (unhappy) is assigned the class *anger* with a strong confidence (i.e. =1) and the class *sadness* with a weak confidence (i.e. =4). The potential disagreement can then be clearly represented together with the degree of likelihood for each discordant judgment.

This annotation scheme offers an advantage of compatibility with current approaches of resolving disagreement. The highest confidence level-0 can be reserved for the project manager to adjudicate on a final decision in case of disagreement, while preserving annotators' various interpretations using a lower confidence level. When the annotation project is carried out via collaborative effort, the "votes" of different annotators can also be shown in terms of the proportion. For example, if the judgments of a group of annotators between class A and class B form a ratio of 8:2, then it can be normalized and represented as <class='A' confidence=1; class='B' confidence=4>.

Furthermore, for the needs of certain tasks such as the training of computational models which requires highly consistent data, the annotations with a low confidence level can be easily filtered out by a confidence threshold (e.g., only the annotated entries with a confidence level-1 or above are included). Hence, our proposal will not be in conflict with existing practices and applications of annotation, while preserving valuable information for the study of interesting linguistic phenomena.

5 Summary

In this paper we address the resolution of inter-annotator disagreement in corpus annotation. While maintaining the importance of consistency criterion, we claim that this does not necessarily mean giving up preservation of multiple interpretations, given that they are plausible and legitimate.

Since ambiguities have rarely been properly recorded in the past annotation projects, we have very limited resources to study them empirically, not to mention the refinement of relevant linguistic theories and/or taxonomies so as to account for and resolve these ambiguities systematically. This has become more and more significant as the interest in annotation in recent decades is moving from the well-studied linguistic systems (e.g. morphology and syntax) towards the under-explored areas (e.g. social acts and emotion). The latter is still at the early stages of development. A solution, we envisage, is to first record the interesting and challenging ambiguous expressions. They are at least as valuable as the linguistic phenomena without disagreement, in terms of providing insights to enrich our understanding towards the understudied linguistic phenomena.

To this end, we suggest an annotation scheme for preserving legitimate disagreement. Despite its rudimentary progress, our scheme is highly compatible with current approaches of disagreement resolution. Consistency can be maintained to cope with the requirements of natural language technology development, while indicating the expressions which are ambiguous and worthwhile for further study.

Acknowledgments

The work is supported by the Dean's Reserve grant sponsored by The Hong Kong Polytechnic University (Project no. 1-ZVB2).

References

Cecilia Ovesdotter Alm. 2010. Characteristics of high agreement affect annotation in text. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW-10)*, pages 118–122.

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

Ying Chen, Sophia Y. M. Lee, and Chu-Ren Huang. 2009. A cognitive-based annotation system for emotion computing. In *Proceedings of the Third Linguistic Annotation Workshop (LAW-09)*, pages 1–9.

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

Yanyan Cui and Ting Chi. 2013. Annotating modal expressions in the Chinese treebank. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM).*

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation - no easy way out! A case from Bangla and Hindi POS labeling tasks. In *Proceedings of the Third Linguistic Annotation Workshop (LAW-09)*, pages 10–18.

Stefanie Dipper and Heike Zinsmeister. 2009. Annotating discourse anaphora. In *Proceedings of the Third Linguistic Annotation Workshop (LAW-09)*, pages 166–169.

Cecily Jill Duffield, Jena D. Hwang, Susan Windisch Brown, Dmitriy Dligach, Sarah E. Vieweg, Jenny Davis, and Martha Palmer. 2007. Criteria for the manual grouping of verb senses. In *Proceedings of the Linguistic Annotation Workshop (LAW-07)*, pages 49–52.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *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*, pages 10–18.

Jisup Hong and Collin F. Baker. 2011. How good is the crowd at "real" WSD? In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW-11)*, pages 30–37.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop (LAW-07)*, pages 132–139.

Youngim Jung and Hyuk-Chul Kwon. 2011. Consistency maintenance in prosodic labeling for reliable prediction of prosodic breaks. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW-11)*, pages 38–46.

Sophia Yat Mei Lee, Ying Chen, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause events: Corpus construction and analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-10)*, pages 1121–1128.

Geoffrey Leech and Elizabeth Eyes. 1997. Syntactic annotation: Treebanks. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 34–52. Addison-Wesley Longman Limited.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn discourse treebank. In *Proceedings of the Language Resources and Evaluation Conference (LREC-04)*.

Jonathan T. Morgan, Meghan Oxley, Emily M. Bender, Liyi Zhu, Varya Gracheva, and Mark Zachry. 2013. Are we there yet?: The development of a corpus annotated for social acts in multilingual online discourse. *Dialogue and Discourse*, 4(2):1–33.

Massimo Poesio and Ron Artstein. 2005. Annotating (anaphoric) ambiguity. In P. Danielsson and M. Wagenmakers, editors, *Proceedings of Corpus Linguistics*, volume 1 of *The Corpus Linguistics Conference Series*.

Changqin Quan and Fuji Ren. 2009. Construction of a blog emotion corpus for Chinese emotional expression analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-09)*, pages 1446–1454.

Jonathon Read and John Carroll. 2012. Annotating expressions of appraisal in English. *Language Resources and Evaluation*, 46(3):421–447.

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2013. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, pages 1–28, April.

Josef Ruppenhofer, Russell Lee-Goldman, Caroline Sporleder, and Roser Morante. 2012. Beyond sentence-level semantic role labeling: Linking argument structures in discourse. *Language Resources and Evaluation*, November.

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

Manfred Stede and Chu-Ren Huang. 2012. Interoperability and reusability: The science of annotation. *Language Resources and Evaluation*, 46(1):91–94.

Yannick Versley. 2006. Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-) reference. In *Proceedings of the ESSLLI Workshop on Ambiguity in Anaphora*.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC-00)*.

Ruifeng Xu, Qin Lu, Kam-Fai Wong, and Wenjie Li. 2007. Annotating Chinese collocations with multi information. In *Proceedings of the Linguistic Annotation Workshop (LAW-07)*, pages 61–68.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated Chinese corpus. In *Proceedings of the 19th International Conference on Computational linguistics (COLING-02)*, pages 1–8.

Annie Zaenen. 2006. Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.