Disambiguating for the Web: A Test of Two Methods

Jonathan Pool
Turing Center, University of Washington
Seattle, Washington, USA
pool@cs.washington.edu

S. M. Colowick
Utilika Foundation
Seattle, Washington, USA
smc@utilika.org

ABSTRACT
The Semantic Web vision expects authors to represent knowledge unambiguously, but their ability and willingness to do so are contested. To evaluate experimentally two disambiguation methods that authors might use, we showed sentences from the Web containing syntactically ambiguous quantification to 386 subjects and asked them to choose between pairs of paraphrasal and/or truth-conditional restatements. The paraphrasal method was mostly superior. Subjects generally found both methods satisfying and were able to achieve substantial consistency and agreement.

Categories and Subject Descriptors
H.1.2 User/Machine Systems – *Human factors, human information processi*ng

H.5.2 User Interfaces – *Natural language*

I.2.4 Knowledge Representation Formalisms and Methods – *Semantic networks*

I.2.6 Learning – *Knowledge acquisition*

I.7.2 Documentation Preparation – *Markup languages*

J.5 Arts and Humanities – *Linguistics*

General Terms
Economics, Experimentation, Human Factors, Languages

Keywords
Ambiguity, Annotation, Disambiguation, Distributed Human Computation, Metadata, Semantic Web

INTRODUCTION
The envisioned Semantic Web would rely on human disambiguation of Web content [3, 6], but this may be unnecessary [4], and even if necessary may be infeasible [5]. Can Web authors improve the retrieval and processing of their documents by preventing ambiguities? If so, how?

To address these questions, we conducted an experiment that gave 386 subjects sentences from the Web containing quantification ambiguities [2]. Subjects disambiguated the sentences with either or both of two methods: paraphrasal selection and truth-conditional selection. We evaluated the methods in terms of subject satisfaction, consistency, speed, and agreement.

METHOD

Participants
Subjects were 200 contractors (paid $0.75 each) recruited through Amazon Mechanical Turk [1] and 186 unpaid volunteers recruited through Internet discussion groups focused on English usage and linguistics. The ability to read and write English was the only prerequisite.

Procedure
We chose 25 English sentences exhibiting quantification ambiguity. Most used the adverbial quantifier "almost always" or "nearly always". For each sentence we identified two interpretations. For each interpretation we wrote two *equivalent* restatements: a paraphrase and a situation description ("truth condition").

Subjects disambiguated each sentence by choosing between the two paraphrases or the two truth conditions, or both. We randomly assigned subjects to four treatment groups: (0) one task per sentence in batches of five, in the order PTPTP (P = paraphrasal, T = truth-conditional); (1) the same as group 0, except in TPTPT order; (2) paraphrasal and truth-conditional tasks together, displayed in that order, for each sentence; and (3) truth-conditional and paraphrasal tasks together, displayed in that order, for each sentence.

After every fifth trial, subjects completed a questionnaire asking how interesting, easy, and useful they felt the study was. Each trial and questionnaire had space for comments.

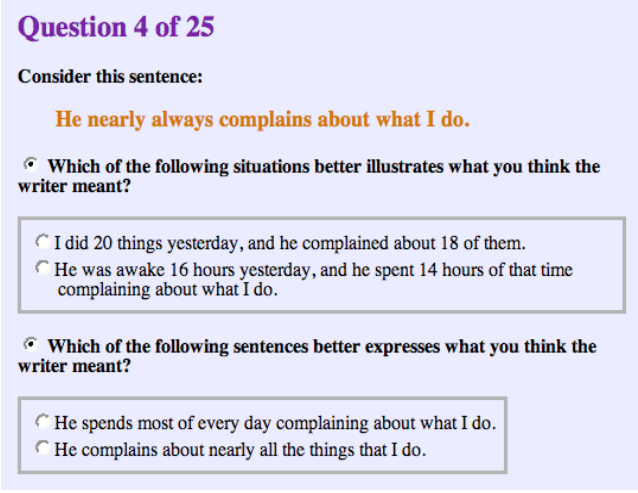

Figure 1. User Interface

The order of the stimulus sentences across trials, and the order of the two alternative responses on each task, were random. Thus, two-task subjects might see a paraphrase and its equivalent truth condition in either corresponding or opposite positions.

We conducted the experiment on the Web, where it was accessible for four days. The instrument may be tested at http://utilika.org/re/aa/. A portion of the user interface is illustrated in Figure 1.

RESULTS

Satisfaction

Subjects expressed moderate satisfaction with the study in all conditions, but tended to rate the study as easiest and most interesting when they had just performed paraphrasal-only disambiguation, and most difficult and most boring when they had just performed both tasks with the truth-conditional task on top. The mean rating of perceived difficulty stayed between 2.2 and 2.3 (on a 0-to-4 scale) throughout the 25 trials.

Another indicator of satisfaction is the 87% completion rate. More one-task subjects finished the study than two-task subjects (90% versus 83%; $p < 0.04$).

Consistency, Speed, and Agreement

The choices made by a two-task subject in a trial were considered *consistent* if the chosen truth condition was equivalent to the chosen paraphrase. In each two-task condition, 82% of the trials were completed consistently. The order in which the tasks appeared on the page did not noticeably affect the consistency rate.

The median time to perform a disambiguation was 20 seconds on one-task trials and 31 seconds on two-task trials. Subjects typically performed paraphrasal disambiguation about 30% faster than truth-conditional disambiguation, possibly because truth-conditional alternatives were longer and more complex and required quantitative reasoning. Disambiguation speed increased with subjects' experience.

The stimulus sentences produced a wide range of inter-subject agreement (from 51% to 94%), which we measured as the fraction of subjects making the majority choice among subjects using the same method on the same sentence. In the aggregate, 77% agreed with the majority. Paraphrasal tasks had larger majorities than truth-conditional tasks (79% versus 75%).

The sentences on which there was greatest agreement tended to be those that subjects disambiguated most rapidly. The degree of agreement among subjects did not substantially increase over the 25 trials.

Subsamples

The unpaid volunteers were slightly less satisfied than paid subjects, but they were more consistent, reached greater agreement, and offered more comments. While doing better, volunteers took more time: Their median trial duration, excluding trials on which comments were added, was 26 seconds, compared with 23 seconds for paid subjects.

DISCUSSION

Our results suggest that humans can disambiguate successfully. After five minutes of practice, subjects applying one method per sentence were resolving ambiguities in 15 to 25 seconds, achieving about 80% inter-method consistency and 80% majority agreement. We consider this sufficient evidence of competence in disambiguation to merit continued investigation.

As subjects repeatedly disambiguated sentences, their speed increased. By the 25th trial, one-task subjects were disambiguating at about 2,200 words per hour, nine times as fast as the 2,000-words-per-day pace of typical human translation [7]. In our subsample analysis, increased speed was associated with greater satisfaction but also with lower quality; better mixtures of these variables might be achieved with training, incentives, and support. This is a topic for future study.

We received numerous comments expressing enjoyment of, and engagement with, the more difficult sentences. Many people appear to like tackling the subtle and intriguing ambiguities that machines have trouble with. This motivation could contribute to the success of active learning strategies in human-machine collaborative disambiguation.

ACKNOWLEDGEMENTS

We are grateful to Emily Bender, Marcus Sammer, and anonymous reviewers for comments on prior versions.

REFERENCES

[1] Amazon.com, Amazon Mechanical Turk (Web site), 2007; http://www.mturk.com/mturk/welcome.

[2] Bach, E., Jelinek, E., Kratzer, A., and Partee, B.H. (eds.), *Quantification in Natural Languages* (Dordrecht: Kluwer, 1995).

[3] Berners-Lee, T., Hendler, J., and Lassila, O., "The Semantic Web", *Sci. Am.,* 284(5), 2001, 34-43; http://www.lassila.org/publications/2001/SciAm.shtml.

[4] Etzioni, O., Banko, M., and Cafarella, M. J., "Machine Reading", 2007 AAAI Spring Symposium on Machine Reading, 2007; http://turing.cs.washington.edu/papers/SS06EtzioniO.pdf.

[5] Marshall, C. C., and Shipman, F. M., "Which Semantic Web?", Hypertext '03 Proceedings, 2003; http://www.csdl.tamu.edu/~marshall/ht03-sw-4.pdf.

[6] Noy, N. F., and McGuinness, D. L., "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford Medical Informatics Technical Report SMI-2001-0880, 2001; http://smi-web.stanford.edu/smi-web/reports/SMI-2001-0880.pdf.

[7] PROZ: The Translators Workplace, "What Is the Realistic Translation Speed?" (Web discussion), 2006; http://www.proz.com/topic/40966.