DP: A Detector for Presuppositions in survey questions

Katja WIEMER-HASTINGS
Psychology Department / Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
kwiemer@latte.memphis.edu

Peter WIEMER-HASTINGS
Human Communication Research Centre
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
peterwh@cogsci.ed.ac.uk

Sonya RAJAN, Art GRAESSER, Roger KREUZ, & Ashish KARNAVAT
Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152
sonyarajan@hotmail.com, graesser@memphis.edu, rkreuz@memphis.edu, akarnavat@hotmail.com

Abstract

This paper describes and evaluates a detector of presuppositions (DP) for survey questions. Incorrect presuppositions can make it difficult to answer a question correctly. Since they can be difficult to detect, DP is a useful tool for questionnaire designer. DP performs well using local characteristics of presuppositions. It reports the presupposition to the survey methodologist who can determine whether the presupposition is valid.

Introduction

Presuppositions are propositions that take some information as given, or as "the logical assumptions underlying utterances" (Dijkstra & de Smedt , 1996, p. 255; for a general overview, see McCawley, 1981). Presupposed information includes <u>state of affairs</u>, such as being married; <u>events</u>, such as a graduation; <u>possessions</u>, such as a house, children, knowledge about something; and others. For example, the question, "when did you graduate from college", presupposes the event that the respondent did in fact graduate from college. The answer options may be ranges of years, such as "between 1970 and 1980". Someone who has never attended college can either not respond at all, or give a random (and false) reply. Thus, incorrect presuppositions cause two problems. First, the question is difficult to answer. Second, assuming that people feel obliged to answer them anyway, their answers present false information. This biases survey statistics, or, in an extreme case, makes them useless.

The detector for presuppositions (DP) is part of the computer tool QUAID (Graesser, Wiemer-Hastings, Kreuz, Wiemer-Hastings & Marquis, in press), which helps survey methodologists design questions that are easy to process. DP detects a presupposition and reports it to the survey methodologist, who can examine if the presupposition is correct. QUAID is a computerized QUEST questionnaire evaluation aid. It is based on QUEST (Graesser & Franklin, 1990), a computational model of the cognitive processes underlying human question answering. QUAID critiques questions with respect to unfamiliar technical terms, vague terms, working memory overload, complex syntax, incorrect presuppositions, and unclear question purpose or category. These problems are a subset of potential problems that have been identified by Graesser, Bommareddy, Swamer, and Golding (1996; see also Graesser, Kennedy, Wiemer-Hastings & Ottati, 1999).

QUAID performs reliably on the first five problem categories. In comparison to these five problems, presupposition detection is even more challenging. For unfamiliar technical terms, for example, QUAID reports words with frequencies below a certain threshold. Such an elegant solution is impossible for presuppositions. Their forms vary widely across presupposition types. Therefore, their detection requires a complex set of rules, carefully tuned to identify a variety of presupposition problems. DP prints out the presuppositions of a question, and relies on the survey methodologist to make the final decision whether the presuppositions are valid.

1 How to detect presuppositions

We conducted a content analysis of questions with presupposition problems to construct a list of indicators for presuppositions. 22 questions containing problematic presuppositions were selected from a corpus of 550 questions, taken from questionnaires provided by the U.S. Census Bureau. The 22 questions were identified based on ratings by three human expert raters. It may seem that this problem is infrequent, but then, these questions are part of commonly used questionnaires that have been designed and revised very thoughtfully.

Additionally, we randomly selected a contrast question sample of 22 questions rated unproblematic with regard to incorrect presuppositions by all three raters. Examples (1) and (2) are questions rated as problematic by at least two raters; examples (3) and (4) present questions that do not contain presuppositions.

> (1) Is that the same place you USUALLY go when you need routine or preventive care, such as a physical examination or check up?
> (2) How much do your parents or parent know about your close friends' parents?
> (3) From date to December 31, did you take one or more trips or outings in the United States, of at least one mile, for the PRIMARY purpose of observing, photographing, or feeding wildlife?
> (4) Are you now on full-time active duty with the armed forces?

Example (1) presupposes the habit of making use of routine / preventive care; (2) presupposes that the respondent has close friends.

As stated above, incorrect presuppositions are infrequent in well-designed questionnaires. For example, questions about details of somebody's marriage are usually preceded by a question establishing the person's marital status.

In spite of this, providing feedback about presuppositions to the survey methodologist is useful. Importantly, QUAID is designed to aid in the design process. Consider a survey on health-related issues. In the context of this topic, a survey methodologist may be interested in how many days of work a person missed because of illness, but not think about whether the person actually has a job. Upon entering the question "how many days of work did you miss last year because of illness" into the QUAID tool, DP would report that the question presupposes employment. The survey methodologist could then insert a question about employment.

Second, there are subtle presuppositions that may go undetected even by a skilled survey designer. These are presuppositions about things that are likely (but not necessarily) true. For example, a question may inquire about a person's close friends (presupposing close friends) or someone's standard place for preventive care (presupposing the habit of making use of preventive care). DP does not know which presuppositions are likely to be valid or invalid, and is therefore more likely to detect such subtle incorrect presuppositions than a human expert.

1.1 The presupposition detector (DP)

We constructed a set of presupposition detection rules based on the content analysis. The rules use a wide range of linguistic information about the input sentences, including particular words (such as "why"), part of speech categories (e.g., wh-pronoun), and complex syntactic subtrees (such as a quantification clause, followed by a noun phrase).

1.1.1 The syntactic analysis component

We used Eric Brill's rule-based word tagger (1992, 1994a, 1994b), the *de facto* state of the art tagging system, to break the questions down into part-of-speech categories. Brill's tagger produces a single lexical category for each word in a sentence by first assigning tags based on the frequency of occurrence of the word in that category, and then applying a set of context-based re-tagging rules. The tagged text was then passed on to Abney's SCOL/CASS system (1996a, 1996b), an extreme bottom-up parser. It is designed to avoid ambiguity problems by applying grammar rules on a level-by-level basis. Each level contains rules that will only fire if they are correct with high probability. Once the parse moves on to a higher level, it will not attempt to apply lower-level rules. In this way, the parser identifies chunks of information, which it can be reasonably certain are connected, even when it cannot create a complete parse of a sentence.

1.1.2 *The presupposition indicators*

The indicators for presuppositions were tested against questions rated as "unproblematic" to eliminate items that failed to discriminate questions with versus without presuppositions. We constructed a second list of indicators that detect questions containing no presuppositions. All indicators are listed in Table 1. These lists are certainly far from complete, but they present a good basis for evaluating of how well presuppositions can be detected by an NLP system. These rules were integrated into a decision tree structure, as illustrated in Figure 1.

Table 1: Indicators of absence or presence of presuppositions

	Presupposition	No presupposition
First word(s)	When VP What time Who VP Why How much How many How often etc. How VP Where V NP	Initial or following comma: - is there - are there Does / do NP have ... Will NP have ... Has / Have NP ... Is / are NP ...
Keywords	usually Possessives: mine, yours, NP's while Indexicals: this, these, such	ever any anybody anything whether if could, would
Specific constructions	V infinitive when NP	

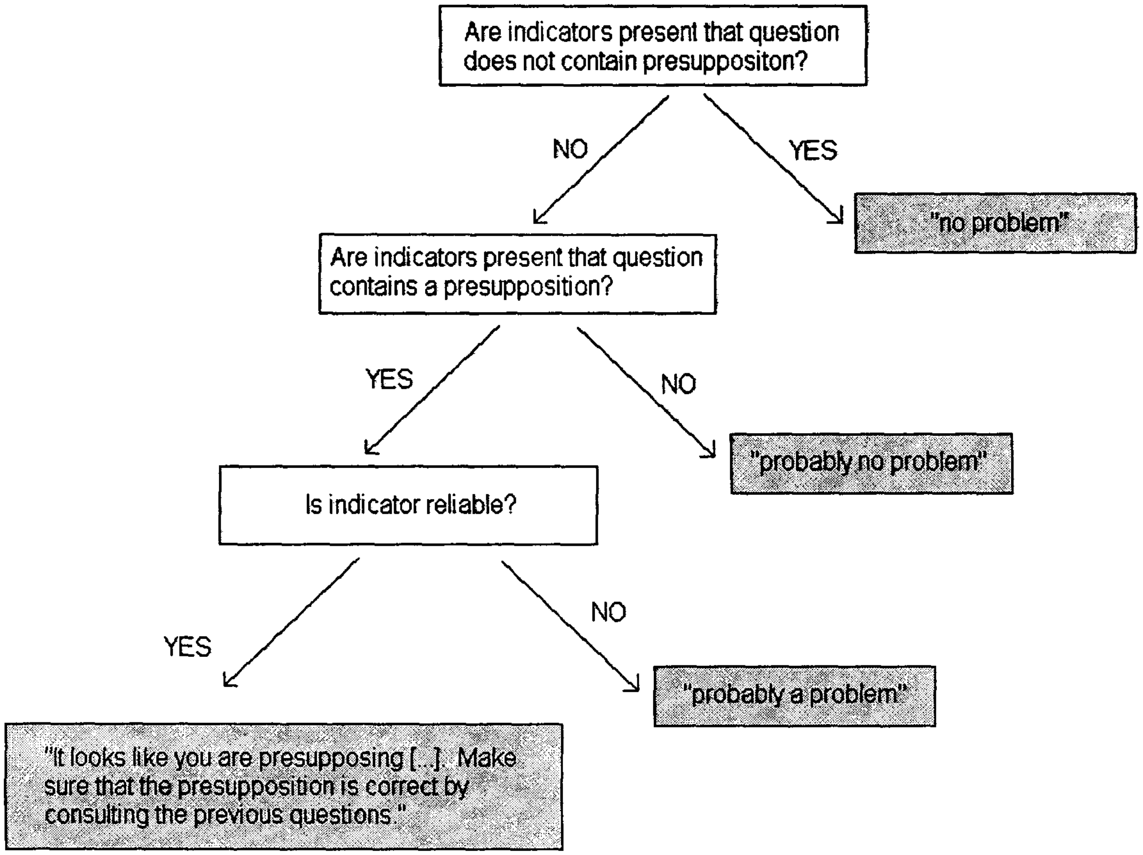
Are indicators present that question does not contain presuppositon?
NO
YES
"no problem"
Are indicators present that question contains a presupposition?
YES
NO
"probably no problem"
Is indicator reliable?
YES
NO
"probably a problem"
"It looks like you are presupposing [...]. Make sure that the presupposition is correct by consulting the previous questions."

Figure 1 : The DP decision structure tree

1.2 Classifying presuppositions

Different types of presuppositions can be distinguished based on particular indicators. Examples for presupposition types, such as events or possessions, were mentioned above. Table 2 presents an exhaustive overview of presupposition types identified in our analysis. Note that some indicators can point to more than one type of presupposition.

Table 2 : Classification of presupposition based on indicators. In the right column, expressions in parentheses identify the presupposed unit.

Indicator	Presupposition type: The question presupposes...
"how often" ...VP "how" aux NP VP "while" ... VP "where" ... VP "why" ... VP	an action (V)
"usually" ... VP "how often", "frequently", etc.	a habit (V)
"how many" NP "where is" NP	an entity: object, state, or person (NP)
Indexicals: "this" / "that" NP "these" / "those" NP "such a(n)" NP	a shared referent or common ground (NP)
"how much" NP ... "how much does" NP "know" "how many" NP ... Possessive pronouns Apostrophe 's': NP's	a possession (NP); exception list: NP's that can be presupposed (name, age, etc.)
"why" S	a state of affairs, fact, or assertion (S)
VP infinitive "why" VP NP	an intention / a goal (infinitive / NP VP)
"who" VP	an agent (A person who VP)
"When" VP ..."when" NP VP	an event (VP)

DP reports when a presupposition is present, and it also indicates the type of presupposition that is made (e.g., a common ground presupposition or the presupposition of a habit) in order to point the question designer to the potential presupposition error. DP uses the expressions in the right column in Table 2, selected in accordance with the indicators, and fills them into the brackets in its output (see Figure 1). For example, given the question "How old is your child?", DP would detect the possessive pronoun "your", and accordingly respond: "It looks like you are presupposing a possession (child). Make sure that the presupposition is correct by consulting the previous questions."

2 Evaluation

In this section, we report summary statistics for the human ratings of our test questions and the measures we computed based on these ratings to evaluate DP's performance.

2.1 Human ratings

We used human ratings as the standard against which to evaluate the performance of DP. Three raters rated about 90 questions from 12 questionnaires provided by the Census Bureau. DP currently does not use context. To have a fair test of its performance, the questions were presented to the human raters out of context, and they were instructed to rate them as isolated questions. Ratings were made on a four-point scale, indicating whether the question contained no presupposition (1), probably contained no presupposition (2), probably contained a presupposition (3), or definitely contained a presupposition (4). We transformed the ratings into Boolean ratings by combining ratings of 1 and 2 ("no problem") versus ratings of 3 and 4 ("problem"). We obtained very similar results for analyses of the ratings based on the four-point and the Boolean scale. For simplicity, we just report the results for the Boolean scale.

2.2 Agreement among the raters

We evaluated the agreement among the raters with three measures: correlations, Cohen's kappa, and percent agreement. Correlations were significant only between two raters (r = 0. 41); the correlations of these two with the third rater produced non-significant correlations, indicating that the third rater may have used a different strategy. The kappa scores, similarly, were significant only for two raters (k = 0.36). In terms of percent agreement, the raters with correlated ratings agreed in 67% of the cases. The percentages of agreement with rater 3 were 57% and 56%, respectively.

DP ratings were significantly correlated with the ratings provided by the two human raters who agreed well ($\underline{r}$ = 0.32 and 0.31), resulting in agreement of ratings in 63% and 66% of the questions. In other words, the agreement of ratings provided by the system and by two human raters is comparable to the highest agreement rate achieved between the human raters.

Some of the human ratings diverged substantially. Therefore, we computed two restrictive measures based on the ratings to evaluate the performance of DP. Both scores are Boolean. The first score is "lenient"; it reports a presupposition only if at least two raters report a presupposition for the question (rating of 3 or 4). We call this measure P_{maj}, a majority-based presupposition count. The second score is strict. It reports a presupposition only if all three raters report a presupposition. This measure is called P_{comp}, a presupposition count based on complete agreement. It results in fewer detected presuppositions overall: P_{comp} reports presuppositions for 29 of the questions (33%), whereas P_{maj} reports 57 (64%).

2.3 Evaluation of the DP

DP ratings were significantly correlated only with P_{comp} (0.35). DP and P_{comp} ratings were in agreement for 67% of the questions. Table 3 lists hit and false alarm rates for DP, separately for P_{maj} and P_{comp}. The hit rate indicates how many of the presuppositions identified by the human ratings were detected by DP. The false alarm rate indicates how often DP reported a presupposition when the human raters did not. The measures look better with respect to the complete agreement criterion, P_{comp}.

Table 3 further lists recall and precision scores. The recall rate indicates how many presuppositions DP detects out of the presuppositions reported by the human rating criterion (computed as hits, divided by the sum of hits and misses). The precision score (computed as hits, divided by the sum of hits and false alarms) measures how many presuppositions reported by DP are actually present, as reported by the human ratings.

Table 3: Performance measures for DP with respect to hits, false alarms, and misses.

	Hit rate	False alarm rate	Recall	Precision	d'
P_{maj}	0.54	0.34	0.66	0.74	0.50
P_{comp}	0.72	0.35	0.72	0.50	0.95

All measures, except for precision, look comparable or better in relation to P_{comp}, including d', which measures the actual power of DP to discriminate questions with and without presuppositions. Of course, picking a criterion with better matches does not improve the system's performance in itself.

3 An updated version of DP

Based on the first results, we made a few modifications and then reevaluated DP. In particular, we added items to the possession exception list based on the new corpus and made some of the no-presupposition rules more specific. As a more drastic change, we updated the decision tree structure so that presupposition indicators overrule indicators against presuppositions, increasing the number of reported presuppositions for cases of conflicting indicators:

```
If there is evidence for a problem, report "Problem"
Else
  if evidence against problem, report "No problem"
  else, report "Probably not a problem"
```

Separate analyses show that the modification of the decision tree accounts for most of the performance improvement.

3.1 Results

Table 4 lists the performance measures for the updated DP. Hit and recall rate increased, but so did the false alarm rate, resulting in a lower precision score. The d' score of the updated system with respect to P_{comp} (1.3) is substantially better. The recall rate for this setting is perfect, i.e., DP did not miss any presuppositions. Since survey methodologists will decide whether the presupposition is really a problem, a higher false alarm rate is preferable to missing out presupposition cases. Thus, the updated DP is an improvement over the first version.

Table 4: Performance measures for the updated DP with respect to hits, false alarms, and misses.

	Hit rate	False alarm rate	Recall	Precision	d'
P_{maj}	0.75	0.44	0.84	0.75	0.8
P_{comp}	0.90	0.52	1.00	0.46	1.3

Conclusion

DP can detect presuppositions, and can thereby reliably help a survey methodologist to eliminate incorrect presuppositions. The results for DP with respect to P_{comp} are comparable to, and in some cases even better than, the results for the other five categories. This is a very good result, since most of the five problems allow for "easy" and "elegant" solutions, whereas DP needs to be adjusted to a variety of problems.

It is interesting that the performance of DP looks so much better when compared to the complete agreement score, P_{comp} than when compared to P_{maj}. Recall that P_{comp} only reports a presupposition if all the raters report one. The high agreement of the raters in these cases can presumably be explained by the salience of the presupposition problem. This indicates that DP makes use of reliable indicators for its performance. Good agreement with the other measure, P_{maj}, would suggest that DP additionally reports presuppositions in cases where humans do not agree that a presupposition is present. The higher agreement with the stricter measure is thus a good result.

DP currently works like the other modules of QUAID: it reports potential problems, but leaves it to the survey methodologist to decide whether to act upon the feedback. As such, DP is a substantial addition to QUAID. A future challenge is to turn DP into a DIP (detector of incorrect presuppositions), that is, to reduce the number of reported presuppositions to those likely to be incorrect. DP currently evaluates all questions independent of context, resulting in frequent detections. For example, 20 questions about "this person" may follow one question that establishes the referent. High-frequency repetitive presupposition reports could easily get annoying.

Is a DIP system feasible? At present, it is difficult for NLP systems to use information from context in the evaluation of a statement. What is required to solve this problem is a mechanism that determines whether a presupposed entity (an object, an activity, an assertion, etc.) has been established as applicable in the previous discourse (e.g., in preceding questions).

The Construction Integration (CI) model by Kintsch (1998) provides a good example for how such reference ambiguity can be resolved. CI uses a semantic network that represents an entity in the discourse focus (such as "this person") through higher activations of its links to other concept nodes. Perhaps models such as the CI model can be integrated into the QUAID model to perform context analyses, in combination with tools like Latent Semantic Analysis (LSA, Landauer & Dumais, 1997), which represents text units as vectors in a high-dimensional semantic space. LSA measures the semantic similarity of text units (such as questions) by computing vector cosines. This feature may make LSA a useful tool in the detection of a previous question that establishes a presupposed entity in a later question.

However, questionnaires differ from connected discourse, such as coherent stories, in aspects that make the present problem rather more difficult. Most importantly, the referent for "this person" may have been established in question number 1, and the current question containing the presupposition "this person" is question number 52. A DIP system would have to handle a flexible amount of context, because the distance between questions establishing the correctness of a presupposition and a question building up on it can vary. On the one hand, one could limit the considered context to, say, three questions and risk missing the critical question. On the other hand, it is computationally expensive to keep the complete previous context in the systems "working memory" to evaluate the few presuppositions which may refer back over a large number of questions. Solving this problem will likely require comparing a variety of different settings.

Acknowledgements

This work was partially supported by the Census Bureau (43-YA-BC-802930) and by a grant from the National Science Foundation (SBR 9720314 and SBR 9977969). We wish to acknowledge three colleagues for rating the questions in our evaluation text corpus, and our collaborator Susan Goldman as well as two anonymous reviewers for helpful comments.

References

Abney, S. (1996a). Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Abney, S. (1996b). Methods and statistical linguistics. In J. Klavans & P. Resnik (Eds.), *The Balancing Act*. Cambridge, MA: MIT Press

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*. ACL.

Brill, E. (1993). *A corpus-based approach to language learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Articial Intelligence*. AAAI Press.

Dijkstra, T., & de Smedt, K. (1996). Computational psycholinguistics. *AI and connectionist models of human language processing*. London: Taylor & Francis.

Graesser, A. C., Bommareddy, S., Swamer, S., & Golding, J. (1996). Integrating questionnaire design with a cognitive computational model of human question answering. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methods of determining cognitive and communicative processes in survey research* (pp. 343-175). San Francisco, CA: Jossey-Bass.

Graesser, A.C., & Franklin, S.P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes, 13,* 279-304.

Graesser, A.C., Kennedy, T., Wiemer-Hastings, P., & Ottati, V. (1999). The use of computational cognitive models to improve questions on surveys and questionnaires. In M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, & R. Tourangeau (Eds.), *Cognition and Survey Research* (pp. 199-216). New York: John Wiley & Sons.

Graesser, A.C., Wiemer-Hastings, K., Kreuz, R., Wiemer-Hastings, P., & Marquis, K. (in press). QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, & Computers.*

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

McCawley, J.D. (1981). *Everything that linguists have always wanted to know about logic.* Chicago: University of Chicago Press.