Capturing a Taxonomy of Failures During Automatic Interpretation of Questions Posed in Natural Language

Shaw-Yi Chaw, James J. Fan[*], Dan G. Tecuci, Peter Z. Yeh[†]
Department of Computer Sciences
University of Texas at Austin
1 University Station C0500
Austin, TX 78712 USA
{jchaw,jfan,tecuci,pzyeh}@cs.utexas.edu

ABSTRACT

An important problem in artificial intelligence is capturing, from natural language, formal representations that can be used by a reasoner to compute an answer. Many researchers have studied this problem by developing algorithms addressing specific phenomena in natural language interpretation, but few have studied (or cataloged) the types of failures associated with this problem. Knowledge of these failures can help researchers by providing a road map of open research problems and help practitioners by providing a checklist of issues to address in order to build systems that can achieve good performance on this problem. In this paper, we present a study – conducted in the context of the Halo Project – cataloging the types of failures that occur when capturing knowledge from natural language. We identified the categories of failures by examining a corpus of questions posed by naïve users to a knowledge based question answering system and empirically demonstrated the generality of our categorizations. We also describe available technologies that can address some of the failures we have identified.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Human Factors

Keywords

NLP, ontology, knowledge based systems, controlled languages, question answering

1. INTRODUCTION

An important problem in Artificial Intelligence (AI) is capturing, from natural language, formal representations rich enough for a knowledge based system to reason over. Solving this problem will enable many tasks, such as allowing naïve users (i.e. users unfamiliar with the underlying knowledge base) to pose questions to a knowledge based reasoner.

Many researchers have studied this problem by developing new algorithms [11, 13, 14, 15] or by integrating existing natural language (NL) and knowledge representation (KR) technologies to build end-to-end systems [1]. However, an equally important, but often overlooked, piece of this problem is characterizing the types of failures that occur when producing rich formal representations from natural language.

Such a characterization is useful to both researchers and practitioners. Researchers can use this characterization as a road-map of open research problems, while practitioners can use it as a checklist of the issues that must be addressed in order to achieve good performance when building systems requiring natural language interpretation. Figure 1 shows an example of a failure which can lead to poor performance if overlooked.

To our knowledge no previous work has studied and cataloged the types of failures that occur when producing formal representations from natural language. A related study on the brittleness of knowledge based systems was conducted by Friedland et al. [10] for the systems developed during the Project Halo [20] pilot study. This study, however, focused only on the types of failures that occur during reasoning as these systems do not have a natural language interface.

[*] Currently at IBM T.J. Watson Research Lab.

[†] Currently at Accenture Technology Labs.

K-CAP'07, October 28–31, 2007, Whistler, British Columbia, Canada.
Copyright 2007 ACM 978-1-59593-643-1/07/0010 ... $5.00

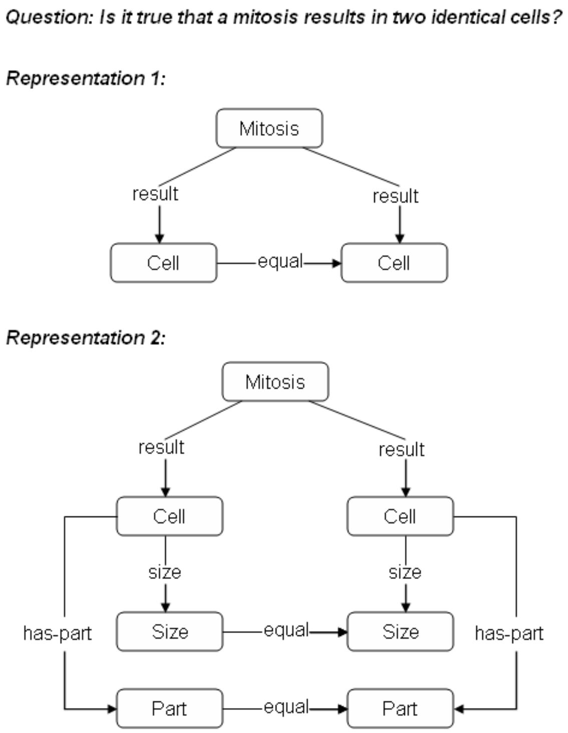


Figure 1: This figure shows an English question and two possible representations. The first representation – although correct – omits details abstracted (e.g. the cells having same size, parts, etc.) using the cue *"identical"*. Hence, this representation may be too shallow for the reasoning requirements of a reasoner and may result in poor performance. The second representation addresses this failure.

In this paper, we present a brittleness taxonomy for the types of failures that frequently occur when capturing, from questions posed in natural language, formal representations that satisfy the reasoning requirements of a reasoner. We first examined a corpus of questions posed by naïve users to build our brittleness taxonomy. We then evaluated the resulting taxonomy on a different corpus to measure the frequency of each failure type. The results showed that the failure types in our taxonomy occur frequently for different question types across three science domains. We also discuss available technologies that can address some of the failures in the taxonomy.

2. BACKGROUND

We conducted our study in the context of Project Halo [9]. The goal of the project is to build a system capable of answering questions posed by naïve users using knowledge bases built by subject matter experts (SMEs) for different science domains. This is especially challenging because the system has to answer questions posed by naïve users who are unfamiliar with the contents and organization of the underlying knowledge base.

One approach to this challenge is to use a template based approach [3, 5, 16, 19] where naïve users pose questions to the knowledge base using a pre-defined set of domain specific question templates. This approach, however, is restrictive and does not allow additional information, such as the question context, to be captured easily. Another approach (and the one taken in Project Halo) is to allow naïve users to pose questions using natural language. Since producing formal representations from natural language is very difficult, controlled languages (CLs) are used. The goal of CL is to avoid difficult problems in natural language processing (e.g. ambiguity or co-reference resolution) by restricting the speaker to a subset of English. CLs have been shown to be robust (and usable) in different research and industrial systems and a CL interpreter called Computer Processable Language (CPL) [4] is used in a system developed for Project Halo.

Figure 2 shows an example question in physics, it's CPL formulation, and its interpretation. Our goal is to catalog the types of failures preventing the underlying knowledge base from answering questions posed by naïve users.

3. TAXONOMY OF FAILURES

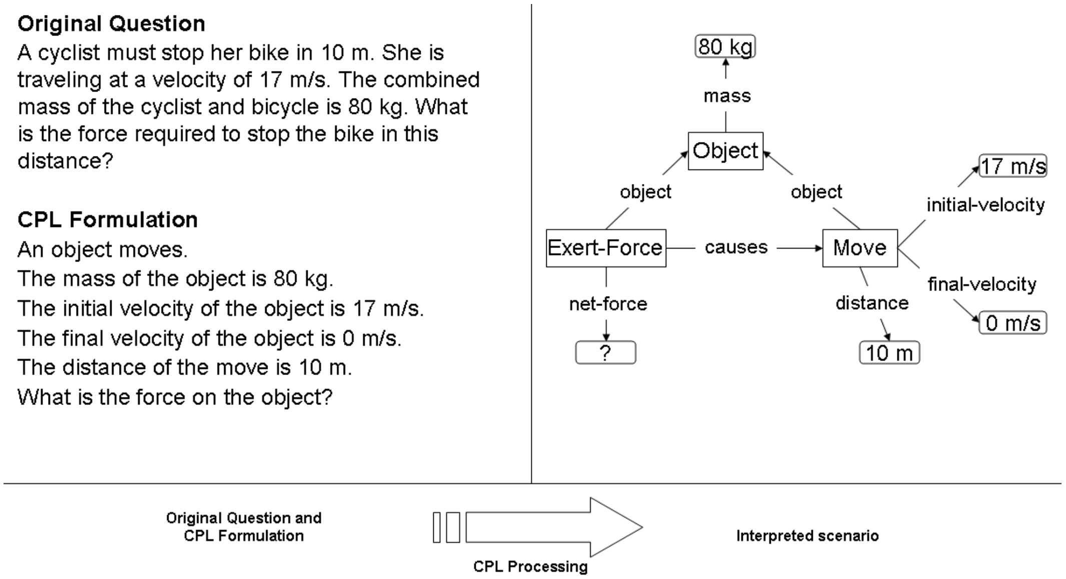
We present a taxonomy of failures that occur when producing formal representations from questions posed in natural language.

3.1 Methodology

Our brittleness taxonomy was built by examining a sample of questions posed by naïve users for two science domains during the Project Halo[20] evaluation conducted in 2006. These questions were posed in CPL by users unfamiliar with either knowledge representation or the contents of the underlying knowledge base. The pilot dataset consists of 50 biology questions and 50 physics questions.

We chose this data set for two reasons. The fact that questions formulated in simplified English continue to be interpreted incorrectly, suggests that these failures are inherent to language. Hence, these are foundational problems that must be addressed before more sophisticated natural language understanding is possible.

This data set is also realistic as it consists of different question types posed by naïve users for two very different science domains. Hence, the types of failures identified in this data set should apply broadly to other domains.



Original Question

A cyclist must stop her bike in 10 m. She is traveling at a velocity of 17 m/s. The combined mass of the cyclist and bicycle is 80 kg. What is the force required to stop the bike in this distance?

CPL Formulation

An object moves.
The mass of the object is 80 kg.
The initial velocity of the object is 17 m/s.
The final velocity of the object is 0 m/s.
The distance of the move is 10 m.
What is the force on the object?

Figure 2: An example illustrating the process of posing and interpreting a question using the CPL controlled language.

Each question and its corresponding formal representation were examined for failures preventing the knowledge base from answering the question.

To guide this process, we focused on failures at the following levels of granularity:

- Word level. We examined the failures associated with interpreting individual words.
- Inter-word level. In addition to individual words, we examined how well the relations between words are interpreted.
- Inter-utterance level. The interpretation of relations between sentences is also prone to failures.
- Overall representation. We also explored the types of failures associated with the overall representation of the input.

These levels provide the right granularity to examine and catalog different sources of failure. The first three levels are well-established. They are the focus of natural language research which has developed algorithms that address specific failures that surface at these levels. We added the "overall representation" level because representations that closely mirror the surface form of an utterance or discourse – which is the focus of most NL research – can still be insufficient for a knowledge based system to reason over.

Recurring failures identified using this methodology were generalized into a separate category in our brittleness taxonomy.

3.2 Brittleness Taxonomy

We present the top-level categories of our brittleness taxonomy in Table 1. Each row in this table lists a category name, a description and an example.

4. EVALUATION

4.1 Experimental Setup and Dataset

We evaluated our brittleness taxonomy by randomly sampling 205 additional question formulations from the 2006 Project Halo evaluation. Aside from biology and physics questions, the evaluation data set also includes questions from the chemistry domain – thus providing a novel domain to evaluate the generality of our taxonomy.

These question formulations were posed by a group of naïve users (four undergraduates and two graduate students). Before the exercise, they underwent six hours of training on how to formulate AP level examination questions using CPL. Subsequently, the users independently posed a set of AP-like exam questions and received answers from the system which used KBs authored by other experts. It is useful to note that AP level questions are often "story problems" that have a question setup (preamble) describing the scenario in which they are to be answered. As shown in Table 2, the percentage of questions containing preambles varies across the domains – 35% for biology, 52% for chemistry and 100% for physics.

	Category	Description	Example
1	**Word Sense Disambiguation**	Mapping the nouns, verbs, adjectives, and adverbs in a sentence to the most appropriate semantic concepts in a knowledge base.	The noun *“top”* can map to the semantic concepts of *Spatial-Region*, *Toy*, etc. Picking the most appropriate concept depends on the context.
2	**Semantic Role Labeling**	Mapping the syntactic relations (e.g. subject, direct object, etc) and prepositional markers in a sentence to the most appropriate semantic relations in a knowledge base.	The prepositional marker *“by”* can map to the semantic relations of *agent*, *instrument*, *caused-by*, etc. Picking the most appropriate semantic relation depends on the context.
3	**Representation Gap**	The knowledge base is missing concepts, axioms, or theories (we call gaps) which prevent information – surfacing in language – that fall into these gaps from being captured and represented.	A knowledge base may **not** have a theory of space (i.e. spatial concepts such as *Region*, *Place*, etc. and spatial relations such as *is-inside*, *encloses*, etc). Hence, spatial information that surfaces in language (e.g. *“The nucleus is inside the cell.”*) cannot be captured and represented.
4	**Implicit Knowledge**		
4.1	**Noun-Noun Compound**	The semantic relationship between noun-noun compounds are always implicit.	The semantic relationship in *“stone lion”* is *material* – i.e. the lion is made of stone.
4.2	**Co-reference**	Co-reference links (both direct and indirect) between multiple expressions within the same sentence (or across different sentences) are always implicit.	In *“John entered the room. He sat down.”*, *“John”* and *“He”* are direct co-references. In *“The man entered the room. The window was closed.”*, *“room”* and *“window”* are indirect co-references.
4.3	**Omission of Assumptions**	Knowledge assumed to be possessed by both the speaker and the listener (e.g. common-sense knowledge or contextual information) are often omitted to facilitate efficient communication.	The sentence *“A cell has a mitochondria.”* omits knowledge of the mitochondria being part of a nucleus and the nucleus being part of the cell.
4.4	**Abstraction**	Some details are abstracted into linguistic cues which must be expanded in order to recover these details.	The sentence *“There are two identical cells.”* abstracts details such as the cells having the same parts, are of same size, etc. into the cue *“identical”*. This cue must be expanded to capture these details in the resulting representation.
5	**Contradiction**	Information in a sentence can contradict existing representations in the knowledge base.	The sentence *“The Kashmir region is part of Pakistan.”* might contract a knowledge base – to reason about geo-political issues – that represents the Kashmir region as a part of India.
6	**Alternative Representations**	The same information can be represented in different ways – each of which is valid.	The sentence *“The cell encloses a nucleus.”* can be represented in two ways. *”encloses”* can be represented as a semantic relation relating the cell to the nucleus. *“encloses”* can also be represented has a reified concept (e.g. a *Be-Contained* event) where the cell and nucleus are the participants.

Table 1: The top-level categories in our brittleness taxonomy. Each row corresponds to a category in our taxonomy, and gives the name of the category, a description, and an example.

Domain	Type 1	Type 2	Type 3	Type 4.1	Type 4.2	Type 4.3	Type 4.4	Type 5	Type 6	None
Biology	42%	45%	27%	38%	14%	14%	9%	1%	25%	9%
Chemistry	47%	26%	13%	42%	35%	2%	40%	0%	0%	20%
Physics	37%	49%	22%	46%	84%	46%	56%	7%	6%	2%
Overall	42%	40%	21%	44%	42%	21%	35%	3%	14%	10%

Table 3: The average frequencies of the different types of brittleness found by the annotators in the data set as percentage of questions in each domain. Because a question may contain more than one type of brittleness, the sum of each row is greater than 100%.

Domain	**Preamble**	**No preamble**	**Total**
Biology	24	44	68
Chemistry	33	30	69
Physics	65	0	68
Total	122	84	205

Table 2: Distribution of questions based on whether they contain a preamble or not.

Three knowledge engineers (KE) were tasked to tag each formulation with all applicable categories from the taxonomy. The KEs were familiar with the categories in the taxonomy. They also had access to the output of the controlled language interpreter, to the answers returned by the system, and to the underlying knowledge-bases used in computing these answers. This information was necessary to appropriately judge category membership.

Altogether, the two KE's tagged 205 formulations each, while the third tagged 100 formulations.

4.2 Results and Discussion

Table 3 shows the result of the evaluation. The Fleiss Kappa statistic [12] for inter-annotator agreement is 0.64, which suggests *substantial agreement* among the KEs. Highlights of the evaluation results are as follows:

- Brittleness is prevalent in the automatic interpretation of questions. In our evaluation data set, only 10% of the questions did not contain any of the brittleness types described in Table 1.

- The brittleness taxonomy captures the most frequently occurring types of brittleness. The nine types of brittleness were found in as high as 84% of the questions rated by the annotators.

- The presence of preambles in questions do not significantly contribute to the amount of brittleness. The biology data set has the fewest number of questions with preambles, yet its brittleness occurrences are comparable to the other domains. This suggests that brittleness is prevalent even in interpreting short questions without a lot of context.

5. PROPOSED SOLUTIONS

Many of the brittleness categories have been studied extensively in isolation, and different approaches have been proposed for specific categories. For example, word sense disambiguation [13], semantic role labeling [11], noun compound interpretation [15] and co-reference resolution [14] are well studied problems in the natural language community. Besides addressing each brittleness category separately, we believe there are inherent, underlying connections among the categories, and it is possible to develop unified solutions addressing multiple categories. We detail some technologies that are useful in overcoming some of the natural language brittleness problems described earlier.

5.1 Flexible Semantic Matching

A semantic matcher takes two representations (encoded in a form similar to conceptual graphs [17]) and uses taxonomic knowledge to find the largest connected subgraph in one representation that is isomorphic to a subgraph in the other. This flexible semantic matcher – as described by Yeh et al. [21] – then uses a library of transformation rules to shift the representations to improve the match. This improvement might enable other subgraphs to match isomorphically, which in turn might enable more transformation rules until the match improves no further.

Flexible semantic matching can address several of the brittleness categories we have identified. For example, Yeh et al. [23] have demonstrated a unified approach for sense disambiguation and semantic role labeling by matching candidate interpretations of a sentence with background knowledge to select the interpretation with the best match. In addition, flexible semantic matching has also been used to address problems of co-reference resolution and omissions by matching the sentences in a discourse with background knowledge to uncover implicit assumptions and to build a coherent semantic representation of what was said [22].

5.2 Interpreting Loose Speak

The Loose-speak interpreter allows naïve users who are unfamiliar with the knowledge-base to effectively pose questions [8]. An occurrence of "loose speak" is a discrepancy between a literal encoding of the natural lan-

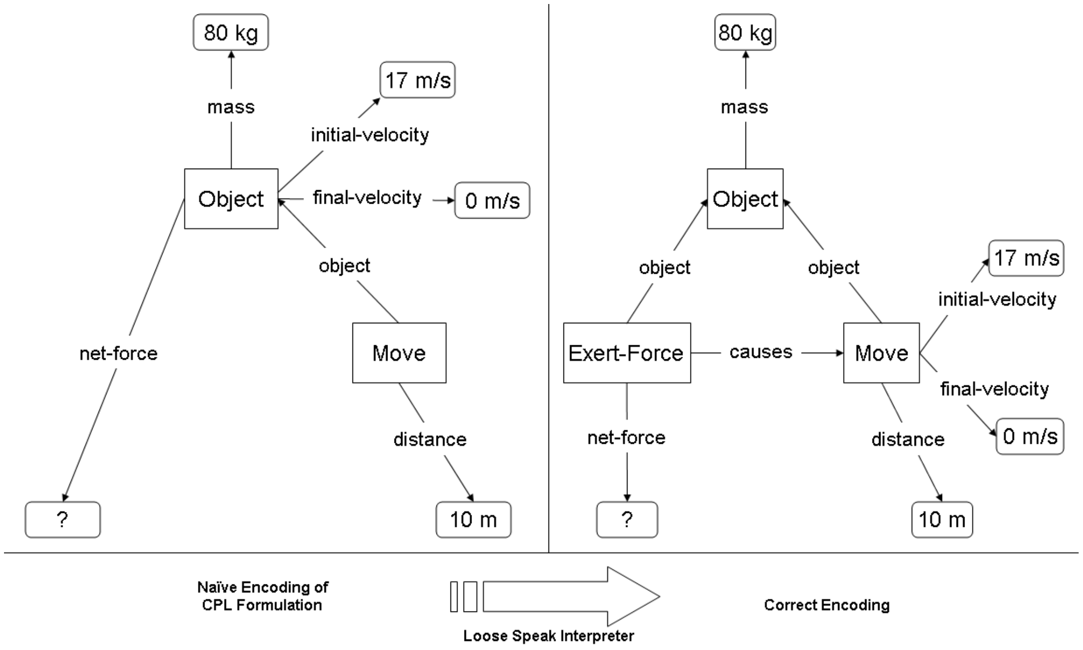


Figure 3: The loose-speak interpreter finds three occurrences of loose speak in the naïve encoding of the controlled language formulation in figure 2. It reassigns the *initial velocity* and the *final velocity* properties to *Move* instead of *Object* based on the slots' constraints. This is an example of type 2 brittleness. It also adds an *Exert-Force* concept as required by the knowledge base. This is an example of type 4.3 Brittleness.

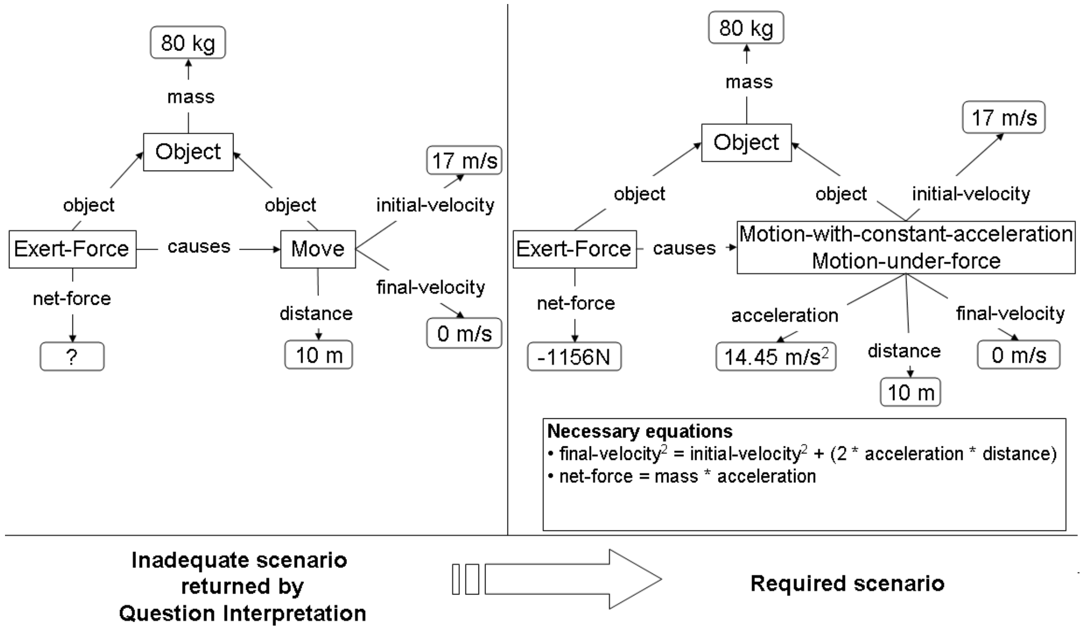


Figure 4: The initial question interpretation (left side) does not contain enough information to solve the problem, i.e. missing equations for calculating the net-force given mass, velocities, and distance. To address this, the problem-solver selects relevant pieces of knowledge in the KB. The right side presents an elaborated scenario containing the necessary equations.

guage input and the representation required by knowledge base. The naïvety of literal encodings of input questions without regard to the idiosyncrasies of the knowledge base is a source of following types of brittleness.

- Words maybe mapped to incorrect slots (as in type 2) because of lack of knowledge of the slot constraints in the knowledge base.
- Noun-noun compounds (type 4.1) may be encoded without explicitly specifying knowledge base specific relations.
- Indirect co-reference links are not explicitly established (type 4.2) in the literal encoding of an utterance, and as a result the utterance encoding does not align with encodings of earlier utterances.
- When assumed information is omitted from the natural language input, the literal encoding is different from what is expected by the knowledge base (type 4.3).

The loose-speak interpreter detects and rephrases discrepancies between naïve and correct encodings by exploring the regularities in different occurrences of "loose speak" to generate encodings that are compatible with the knowledge base. Figure 3 shows the loose-speak interpreter correcting the naïve encoding of the CPL formulation in Figure 2.

The loose-speak interpreter has been shown to be effective in detecting and correcting the specific brittleness types mentioned above [6, 7, 8], however, it does not address other failures in the brittleness taxonomy.

5.3 Problem Solving

Questions posed to knowledge based systems typically omit information that is essential for the reasoner to generate an answer. This incompleteness is due to the nature of question asking and the brevity of human communication. Problem solvers, therefore, take on the responsibility of relating facts in a question with relevant pieces of knowledge in the knowledge base before the question can be answered. Automating this process is especially challenging when the problem solver is meant to work with a variety of knowledge bases. Chaw et al. [2] have studied this problem in the context of a system used to answer questions posed by naïve users who are unfamiliar with the content and organization of the knowledge bases. This separation between the builders and the users of the knowledge bases requires the problem solving system to elaborate the inadequate interpretations automatically with information drawn from the knowledge base being queried. Figure 4 shows the elaboration of the inadequate interpretation for the question in Figure 2.

5.4 Episodic Memory

Humans make use of past experiences to improve both performance and competence. For example, students often prepare for tests by solving practice questions for a given syllabus. The experience in interpreting and solving these questions provides guidance when new questions are attempted by remembering what worked in the past and what did not. Tasks such as flexible semantic matching, interpreting loose-speak, and problem solving are computationally expensive procedures that are aggravated by the complexity and size of the knowledge base. We believe leveraging past experiences can offer performance and usability improvements in question answering applications. A generic episodic memory that can be easily integrated to aid interpretation and problem solving is described in [18].

6. CONCLUSION

In this paper, we presented a study cataloging the types of failures – we call a brittleness taxonomy – that occur when capturing, from natural language, formal representations that satisfy the reasoning requirements of a reasoner. We conducted this study in the context of Project Halo by examining a corpus of questions posed by naïve users. We presented the resulting taxonomy and evaluated it on a different corpus of questions to examine the frequency of each failure type.

We found the failure types in our taxonomy to occur frequently in questions posed by naïve users for three different science domains. These results are encouraging as they show the generality of our brittleness taxonomy.

We also described technologies that can address the various types of failures in our taxonomy and provided a description for each one.

We hope that others can benefit from and build upon our work in the following ways. Ideally, researchers can build upon our work by either extending and refining the categories we have identified or using our taxonomy as a road-map to identify research problems. Also, practitioners can benefit from our work by using our taxonomy as a checklist of issues to address in order to achieve good performance when building systems that require natural language interpretation.

7. ACKNOWLEDGMENTS

We are grateful to Vulcan Inc for funding this work as part of Project Halo.

8. REFERENCES

[1] K. Barker, B. Agashe, S. Chaw, J. Fan, M. Glass, J. Hobbs, E. Hovy, D. Israel, D. S. Kim, R. Mulkar, S. Patwardhan, B. Porter, D. Tecuci, and P. Yeh. Learning by Reading: A Prototype System, Performance Baseline and Lessons

Learned. In *Proceedings of Twenty-Second National Conference on Artificial Intelligence*, 2007.

[2] S. Chaw, K. Barker, B. Porter, and P. Yeh. An Ontology-Independent Problem Solver. Technical report, University of Texas at Austin, May 2007.

[3] P. Clark, K. Barker, B. Porter, V. Chaudhri, S. Mishra, and J. Thomere. Enabling domain experts to convey questions to a machine: A modified, template-based approach. In *Proceedings of Second International Conference on Knowledge Capture*, 2003.

[4] P. Clark, S. Chaw, K. Barker, V. Chaudhri, P. Harrison, J. Fan, B. John, B. Porter, A. Spaulding, J. Thompson, and P. Z. Yeh. Capturing and Answering Questions Posed to a Knowledge-Based System. In *Proceedings of Fourth International Conference on Knowledge Capture*, 2007.

[5] P. R. Cohen, R. Schrag, E. K. Jones, A. Pease, A. Lin, B. Starr, D. Gunning, and M. Burke. The DARPA High-Performance Knowledge Bases project. *AI Magazine*, 19(4):25–49, 1998.

[6] J. Fan, K. Barker, and B. Porter. The knowledge required to interpret noun compounds. In *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence*, 2003.

[7] J. Fan, K. Barker, and B. Porter. Indirect anaphora resolution as semantic path search. In *Proceedings of Third International Conference on Knowledge Capture*, 2005.

[8] J. Fan and B. Porter. Interpret loosely encoded questions. In *Proceedings of Nineteenth National Conference on Artificial Intelligence*. AAAI Press, 2004.

[9] N. Friedland, P. Allen, P. Matthews, M. Witbrock, D. Baxter, J. Curtis, B. Shepard, P. Miraglia, J. Angele, S. Staab, E. Moench, H. Oppermann, D. Wenke, D. Israel, V. Chaudhri, B. Porter, K. Barker, J. Fan, S. Chaw, P. Yeh, D. Tecuci, and P. Clark. Project Halo: Towards a Digital Aristotle. *AI Magazine*, 2004.

[10] N. Friedland, P. Allen, M. Witbrock, J. Angele, S. Staab, D. Israel, V. Chaudhri, B. Porter, K. Barker, and P. Clark. Towards a quantitative, platform-independent analysis of knowledge systems. In *Proceedings of Ninth International Conference on the Principles of Knowledge Representation and Reasoning*, 2004.

[11] D. Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3), 2002.

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

[13] R. Mihalcea, P. Tarau, and E. Figa. PageRank on Semantic Networks, with Application to Word Sense Disambiguation. In *COLING*, 2004.

[14] R. Mitkov. *Anaphora Resolution*. Longman, London, 2002.

[15] V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz. Learning noun-modifer semantic relations with corpus-based and wordnet-based features. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI '07)*, 2006.

[16] R. Schrag, M. Pool, V. Chaudhri, R. C. Kahlert, J. Powers, P. Cohen, J. Fitzgerald, and M. S. Experimental evaluation of subject matter expert-oriented. Technical report, Information Extraction and Transport, Inc., August 2002. Proceedings of the 2002 PerMIS Workshop, August 13-15, 2002, NIST Special Publication 990, pp. 272-279.

[17] J. F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984.

[18] D. Tecuci and B. Porter. A Generic Memory Module for Events. In *Proceedings to the FLAIRS20 Conference*, Key West, FL, 2007.

[19] Teknowledge Corporation. Rapid Knowledge Formation project, 2002. http://reliant.teknowledge.com/RKF.

[20] Vulcan Inc. Project Halo, 2003. http://projecthalo.com.

[21] P. Yeh, B. Porter, and K. Barker. Using Transformations to Improve Semantic Matching. In *KCAP*, 2003.

[22] P. Yeh, B. Porter, and K. Barker. Matching Utterances to Rich Knowledge Structures to Acquire a Model of the Speaker's Goal. In *KCAP*, 2005.

[23] P. Yeh, B. Porter, and K. Barker. A unified knowledge based approach for sense disambiguation and semantic role labeling. In *Proceedings of Twenty-First National Conference on Artificial Intelligence*. AAAI Press, 2006.