"Out of the Box" Information Extraction: a Case Study using Bio-Medical Texts

Niranjan Balasubramanian, Stephen Soderland, Mausam, Oren Etzioni, and Robert Bart
University at Washington

ABSTRACT

A major obstacle to Information Extraction (IE) from text is the knowledge engineering required for each new domain: specifying the relations of interest, manually encoding extraction rules, or selecting and hand-tagging the training examples necessary to automatically learn such rules. The history of IE is one of increasing automation, and recent systems have reduced manual effort to "50 person hours" [8].

This paper presents the first IE system tested on new domains "out of the box"—without *any* manual effort. The paper reports on a case study assessing the performance of the ReVerb Open IE system on a sample of MedLine abstracts and on a biology textbook.

To date, ReVerb has only been tested on massive, highly redundant Web corpora. Since scientific language is more complex than typical Web text, and since our bio-medical texts exhibit essentially no redundancy, we expected *a priori* ReVerb to yield modest precision. In fact, ReVerb achieves precision 0.9 (at recall of 0.28) on MedLine, and precision 0.9 (at recall of 0.14) on the biology textbook; biology recall climbs to 0.57 at precision 0.7.

Error analysis revealed two main sources of error – conditionally true statements and n-ary relations. Improving ReVerb to identify these cases doubled recall (at precision 0.9) on the biology textbook, and yielded small gains on our other corpora as well.

1. INTRODUCTION AND MOTIVATION

Information Extraction (IE) systems seek to glean factual assertions from unstructured text. The history of IE is one of increasing automation [4]. The first IE systems relied on hand-crafted, domain-specific rules (*e.g.*, [1]). The next generation of IE systems, beginning with [24, 19], automatically learned extractors from hand-tagged sentences. However, the creation of suitable training data for IE still requires substantial effort and expertise. Moreover, the amount of manual effort scales linearly with the number of relations of interest, and these target relations must be specified in advance. The KnowItAll Web IE system [5] took the next step forward by automatically labeling its own training examples using a small set of domain-independent extraction patterns. This approach led to *Open IE* [2] which introduced a simple linguistic theory that, coupled with the massive redundancy of the Web, enabled Open IE systems to dispense with both pre-specified relations and hand-labeled examples for each relation.

Table 1: Examples of Open IE tuples extracted from MedLine abstracts and a biology textbook.

Arg1	Relation	Arg2
implant motility	was significantly better with	the myoconjunctival technique
early rises in histamine	were temporally related to	the immediate erythema
peripheral proteins	are not embedded in	the lipid bilayer
fishes that live in extreme cold	have membranes with	a high proportion of unsaturated hydrocarbon tails

Recently, there has been growing interest in "extreme extraction" scenarios where researchers are challenged to field a working extraction system for a particular corpus in limited time ("50 person hours of work" [8]). Extreme extraction is helpful when a new corpus suddenly becomes available and extraction is urgent (*e.g.*, Bin Laden's laptop); in addition, extreme extraction limits the *cost* of developing an IE system for a particular corpus. In 50 person hours of work, Freedman *et al.* achieved a precision of 0.52 (at recall of 0.49) for five relations. Our biology textbook consists nearly 25,000 sentences where the number of relations is estimated to be upwards of 5,000[1]. Moreover, the set of relations of interest is unknown in advance. Clearly, a more scalable approach is necessary in this case.

Could Open IE be utilized as the basis for a kind of extreme extraction? If successful, Open IE could enable "out of the box" extraction on a new corpus without delay or manual effort.[2] We investigates this question via a case study using a sample of MedLine abstracts and the text of a college biology textbook [18].

While Open IE systems have achieved high *precision* on massive Web corpora, Open IE has not been applied to smaller, domain-specific corpora that exhibit little redun-

[1]Estimate computed using normalized relation strings.

[2]The Open IE task definition differs somewhat from that of the traditional extraction. We consider this issue in depth in the next section.

dancy. In contrast to the Web, strong *recall* is essential in smaller corpora to avoid missing key information. Moreover, current Open IE systems extract binary relational tuples of the form (Arg1, Relation, Arg2) such as (*Mitt Romney, debated, multiple Republican challengers*). It is unclear, *a priori*, how much information is expressed in this form in biomedical corpora. Finally, compared to "typical" Web text, the bio-medical texts include specialized scientific vocabulary, and complex technical sentences that could hurt both the precision and recall of Open IE.

In our experimental case study, we applied the ReVerb system [7, 6] to sentences drawn from MedLine abstracts and from the biology textbook [18] used in the HALO project [9]. Table 1 shows examples of the relational tuples extracted.

We make the following main contributions:

- We present the first experimental report on the application of an Open IE system "out of the box" to a small, focused corpora. We find that at precision 0.9, ReVerb achieved recall of 0.28 for MedLine abstracts and a recall 0.14 for the biology textbook; at precision 0.7, ReVerb recall climbed to 0.57 on the biology text. ReVerb's performance on these specialized scientific texts are comparable to its performance on general web texts.
- Our out-of-the-box results, led us to identify two weaknesses of ReVerb: conditionally true sentence constructions and n-ary relations.
- In response, we improved ReVerb's confidence function, which doubled recall at precision 0.9 on the biology text, and also yielded small gains on MedLine abstracts and Web text, demonstrating the generality of the improvements.

The remainder of this paper is organized as follows. We describe the ReVerb Open IE system and evaluate its out-of-the-box performance on MedLine abstracts and the biology textbook. We then describe improvements to ReVerb to identify conditionally true and n-ary extractions and evaluate the improved system on MedLine, the biology textbook, and general Web text. To demonstrate the utility of ReVerb's extractions, we evaluate it on the task of generating questions for textbook study guide. We then discuss related work and present our conclusions.

2. OPEN INFORMATION EXTRACTION

Conventional approaches to Information Extraction rely on a pre-specified set of relations and a set of hand-tagged training examples for each relation. A relation-specific extractor is then trained based on this training data. However, this approach is not feasible when extractions are needed immediately in a new domain or when the set of relations of interest is not known *a priori*. Moreover, these approaches do not scale well to large sets of relations, since manual annotations are required for each relation.

The Open IE paradigm [2, 4] overcomes these challenges by automatically identifying each *relation phrase* in a sentence, and the phrase's corresponding arguments. For instance, given the sentence, "McCain fought against Obama, but finally lost the 2008 election," the ReVerb Open IE system extracts two tuples, (*McCain, fought against, Obama*), and (*McCain, lost, the 2008 election*).

The automatic identification of relation phrases enables the extraction of *arbitrary* relations from sentences, obviating the restriction to a pre-specified vocabulary. This property enables Open IE to apply out-of-the-box to any English corpus. Of course, Open IE extractions do require further processing to be mapped to an ontology as shown in [25]. Yet, a number of papers have shown that, even without ontological mapping, Open IE extractions can yield common sense knowledge in the form of selectional preferences [20], functional relations [15], and inference rules [22, 3]. Moreover, Open IE extraction are valuable to people wanting to analyze the entities and relations mentioned in a corpus. See, for example, ReVerb's extractions from a 500 million Web page corpus at: `openie.cs.washington.edu`. Finally, this paper shows how Open IE can support the generation of questions for students interacting with digital versions of textbooks.

Several approaches have been proposed for Open IE, *e.g.*, TextRunner [2], WOE [29], and StatSnowBall [31]. Closely related approaches include *Preemptive IE*, which clusters documents and then parse subtrees to find common extraction patterns [23], and distant supervision [16, 12], which is capable of learning to extract large numbers of relations.

In this paper, we utilize ReVerb2.0, which was recently shown to more than double the area under the precision-recall curve compared to previous methods [6].[3] For brevity, we omit the version number "2.0" throughout the paper.

ReVerb processes a sentence in three steps – relation phrase extraction, argument identification, and extraction ranking. For relation phrase extraction, ReVerb utilizes a general constraint on the syntactic form of relation phrases, which is encoded as a sequence of part-of-speech patterns: verb, verb followed by a preposition, verb followed by a noun phrase then preposition *etc.* [7]. For massive corpora, ReVerb implements an additional constraint – after extracting all potential relation phrases in a sentence, it checks for the number of distinct argument pairs that appear with each candidate relation phrase in the corpus. ReVerb retain only relation phrases that occur with more than k distinct argument pairs. This constraint is not invoked when ReVerb is run on small corpora.

ReVerb utilizes ArgLearner [6] for argument identification. ArgLearner invokes CRF-based classifiers to detect left and right boundaries of the two arguments. The classifiers' features are based on lexico-POS patterns that indicate various linguistic constructions common in text, such as relative clauses, compound verbs, and prepositional attachments.

As a final step, ReVerb assesses the correctness of the extracted tuple with a logistic regression classifier. The features are standard NLP features such as sentence length, POS tags of the predicate, context immediately surrounding the tuple, and whether arguments are proper nouns. See [7] and ReVerb's online documentation for a complete description of ReVerb. The training data for learning this classifier is obtained by judging a set of randomly sampled relation tuples extracted from web sentences. Note that this training data creation is not relation specific and does not affect the out-of-the-box application of ReVerb.

Open IE systems such as ReVerb have been shown to achieve good performance on massive Web corpora. How-

[3] ReVerb 2.0 is available at *http://reverb.cs.washington.edu*

ever, it was unclear *a priori* how they would perform on far smaller bio-medical texts. Since there is little redundancy in smaller corpora, recall becomes much more important. Moreover, scientific texts use specialized vocabularies and more complex grammatical constructions, compared with typical Web text. The following section presents the results of our case study.

3. EVALUATING OPEN IE "OUT OF THE BOX"

We evaluated ReVerb's performance on a sample of Medline abstracts[4] and on a survey textbook for first year biology. Our evaluation attempts to answer the following questions: (1) Is ReVerb suitable for out-of-the-box extraction form scientific texts? (2) Does ReVerb extract information that is *useful* to a domain-expert? (3) Does the out-of-the-box application offer novel insights to improve ReVerb's Open IE in general?

We measure the performance of ReVerb using manual annotations. Two computational linguists (one of them an author of this paper), provided judgments on the correctness of the system extractions. An extraction was tagged correct if the sentence asserts the relation to be true and if the following conditions were met: 1) Its relation phrase represented a valid relation between its arguments, and 2) the arguments are not truncated incorrectly and do not include extraneous information. The annotators agreed more than 77% of the cases for correctness tagging. We also used domain experts to create a gold standard of possible relational tuples by manually extracting relational tuples from each sentence.

MedLine Abstracts - There has been considerable interest in automatic extraction from MedLine journal abstracts as an aid in keeping up with the explosion of bio-medical research. In this work, we target extraction from sentences that contain medical conclusions. We use a random sample of 200 sentences that were provided to us by an independent domain expert who used keyword search to find conclusion statements for his own project.

We processed these sentences using ReVerb with parameters set for high recall (*i.e.*, no lexical constraint), but no domain-specific training or knowledge engineering. The ReVerb extractions were judged for correctness by the annotators and the manual extractions, both binary and N-ary, were used to estimate the total number of possible relations.

Table 2 shows summary statistics of the test set. We obtain a total of 358 possible relational tuples from 200 sentences, 218 of them binary (61% of the tuples). ReVerb extracted 166 correct binary tuples, 76% of the possible binary relation tuples and nearly half of all relation tuples.

Figure 1, shows recall and precision using ReVerb's built-in confidence function to rank extractions. ReVerb obtained recall of 0.29 at precision 0.9 for binary relations in MedLine abstracts, and recall 0.37 at precision 0.8. These results coupled with the steady increase in precision for increasing confidence thresholds suggest that ReVerb's confidence function, which was trained on Web text generalizes to bio-medical texts.

Biology Textbook - Extracting useful information from digital text books can help build tools to assist students in learning. We target extraction from a biology text book [18] used in the HALO book project [9]. We created a test set of 200 sentences from two randomly chosen chapters of the textbook and processed them with ReVerb as above. For this corpus we used four domain experts (biology undergrad students) as annotators in addition to the two computational linguists. The linguists annotated the corpus exhaustively, manually extracting all possible binary and n-ary tuples, and also tagged ReVerb extractions using the correctness criteria mentioned earlier. The biology students also tagged the extractions as *useful*, *i.e.* whether they represent a stand-alone useful biological relation in the context of the textbook. We created our test set by reconciling the tags for possible extractions, correctness, and usefulness.

Table 2 shows summary statistics for this test set as well. There were a comparable number of possible binary relations as in MedLine, but more n-ary relations. Correct extractions by ReVerb account for a similar proportion of the binary relations (83%) and over half of all relations. In particular, ReVerb finds the bulk of useful binary relations, over 78% of the useful binary relations and 55% of all useful relations including n-ary.

Recall and precision for the biology textbook are shown in Figure 1. Here recall is 0.14 at precision 0.9; and 0.57 at precision 0.70. Even at its maximum recall of 0.83, ReVerb's precision is above 0.60. As with the MedLine results, we find that the precision-recall curve for the biology text shows the generalizability of ReVerb's confidence function. As we will see in the next section, improvements to Re-

Table 2: Statistics on the labeled test set of MedLine and Biology Text Book extractions.

MedLine	All Possible	ReVerb
Binary	218	166 (76%)
N-ary	140	-
Total	358	166 (46%)
Bio textbook	**All Possible**	**ReVerb**
Binary	189	157 (83%)
N-ary	65	-
Total	254	157 (62%)

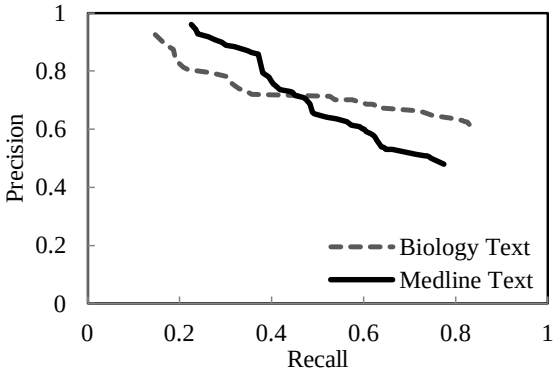

Figure 1: ReVerb out-of-the-box achieves recall 0.29, at precision 0.9 for MedLine abstracts. For a biology text, recall is 0.14 at precision 0.9, and 0.57 at precision 0.70.

[4]http://pubmed.gov

VERB's confidence metric can double the recall at 0.9.

Error Analysis - These results demonstrate that an out-of-the-box application of REVERB, which was developed and tested on Web sentences, performs well on a new specialized domain such as biology. While this is a positive result, the out-of-the-box application is not without errors. We analyze the tuples that were marked as incorrect by the annotators and find two major sources of errors.

1. Conditionally True Relations – Technical information in biology is sometimes specified as relationships that are valid only under a certain condition or context. This accounts for 26% of the errors in MedLine abstracts and 13% of the errors in the biology textbook. REVERB's extraction patterns and confidence metric are based primarily on local context, and have no mechanism for detecting such contextual constructions. Here are examples of this.

   A. Two Dutch scientists reasoned that cell membranes must be phospholipid bilayers.
   B. Other studies have found that equivalent efficacy is reached at lower doses of HFA-BDP than CFC-BDP.
   C. In animal cells, the assembly of spindle microtubules starts at the centrosome.

   The extraction (cell membranes, must be, phospholipid bilayers) from sentence A and the extraction (equivalent efficacy, is reached at, lower doses of HFA-BDP than CFC-BDP) from sentence B are not factual assertions. The extraction (the assembly of spindle microtubules, starts at, the centrosome) from sentence C is only asserted to be true in animal cells.

2. N-ary/Nested Relations – Since REVERB extracts only binary relations, it misses out on relations that require more than two arguments. These can be from verbal relations that require multiple arguments, or from nested sentence structures. Creating a binary tuple for an n-ary or nested relation accounts for 10% of the errors on MedLine and 12% of the errors on the biology textbook. Again, here are examples from MedLine and the biology textbook.

   D. This molecular arrangement provides their hydrophobic parts with a nonaqueous environment.
   E. This will spare children and their families the trauma and pain of cannulation.
   F. This study provides evidence that HD IL-2 should remain the preferred therapy for patients with metastatic renal cell carcinoma.

   The extraction (This molecular arrangement, provides, their hydrophobic parts) from sentence D is incomplete – "provide" is not a binary relation. The extraction (This, will spare, children and their families) from sentence E is also incomplete, missing what they are spared. REVERB might extract the incomplete (This study, provides, evidence) from sentence F, or the non-factual (HD IL-2, should remain, the preferred therapy for patients with metastatic renal cell carcinoma).

Table 3: Features added to the ReVerb confidence classifier to identify conditionally true and nested relations.

Conditionally True / Hypothetical Relations:
hypothetical indicator before arg1
if immediately before arg1
that,which,who immediately before arg1
modal verb before arg1
that,which,who between arg1 and pred
communication verb before arg1
communication verb after arg2
cognition verb before arg1
N-ary or Nested Relations:
comma immediately before arg1
NP immediately after arg2
non-period punct immediately after arg2
that after arg2
verb in arg2
head of relation phrase is cognition verb
head of relation phrase is communication verb
head of relation phrase is n-ary verb

In summary, while REVERB achieves good out-of-the-box precision and recall of binary relations, it fails to handle complex relationships. REVERB was designed to capture simple binary relationships on Web texts, where redundancy of information can help identify the most salient information. However, in a closed corpus such as MedLine abstracts or our biology textbook where relationships generally expressed only once, recall is vital, and handling complex relations becomes more important. In the next section, we design improvements to ReVerb's confidence function to address the Conditionally True and N-ary relations.

4. IMPROVEMENTS TO THE CONFIDENCE FUNCTION

Analysis of REVERB errors on biology text shows the importance of looking beyond local context in assessing the correctness of extracted tuples, a problem common to other genres with complex sentence structures. REVERB's current confidence function [7] is designed to judge whether the predicate is truly a relation between its arguments. To this effect, it uses primarily local, surface-level features in and around the relation phrases and arguments. These include sentence length, coverage of an extraction in a sentence, the part-of-speech labels of a relation and of the immediate context of the arguments.

However, these features are not effective in identifying cases where the extraction may be an accurate local substructure of the sentence, but yet not express a factual assertion (conditionally-true). Moreover, these features also do not help identify n-ary relations.

We introduce novel features for REVERB's confidence classifier to test for these two cases. We use a separate development set for creating these features. We add the new features to REVERB's classifier and in keeping with the Open IE spirit, we train this extended classifier on a training set of general Web text.

Conditionally True Relations: We seek features that indicate when the extraction may not be explicitly asserted

by the sentence. We implement features shown in Figure 3, including features based on lists of terms we mined from the development set of Web text.

The first features indicate the presence of a hypothetical lexical item {if, whether, though, although, suppose} before Arg1. This helps identify hypothetical or supposition statements. A related feature identifies modal verbs {may, might, would, could, should} before Arg1, since they also suggest non-assertions. Other features looks for a communication verb, {deny, declare, promise, ...}, or a cognition verb, {think, believe, realize, ...} before Arg1.

N-ary/Nested Relations: These are cases where a binary tuple cannot represent the relation or where a nested tuple structure is needed – binary extractions leave out essential information and are necessarily incorrect. Figure 3 shows features that help identify such binary tuples, which include verbs often occurring in n-ary relations {give, put, send, *etc.*} found as head of the relation.

While these features help with nested extractions considerably, identifying general n-ary relations is still a challenge. Often additional arguments are expressed via prepositional phrases (PP) after arg2. Identifying these requires us to solve the PP attachment problem, which is notoriously hard.

We add these new features to ReVerb's original features. Using this extended feature set, we retrain the confidence function on a training set of 1,000 Web extractions. We now present the experimental results of using the new confidence function on both the Web and as bio-medical text.

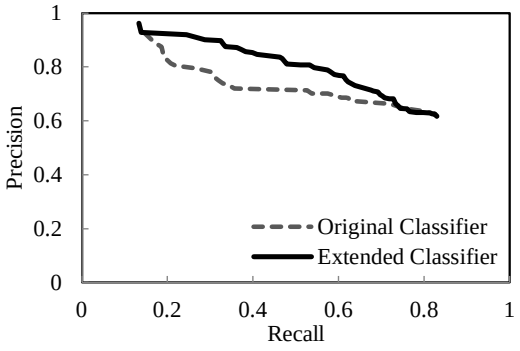

Figure 2: On the biology textbook, the new features more than double recall from 0.14 to 0.29 at precision 0.9, and from recall 0.22 to 0.53 at precision 0.8.

4.1 Results of Improved Confidence

We compare ReVerb's original confidence function and our improved confidence function on the biology textbook, the MedLine abstracts, and a test set of 250 extractions sampled from a Web-based corpus.

As Figure 2 shows, the features to identify conditionally true and n-ary relations give a dramatic improvement on the biology textbook, particularly at high precision. At precision 0.9 the new confidence function boosts recall from 0.14 to 0.29, and at precision 0.8 from recall 0.22 to 0.53.

We found a more modest improvement on MedLine, where recall increased from 0.37 to 0.39 at precision 0.8 and a larger boost from 0.47 to 0.59 at precision 0.7. However, the new classifier did not raise precision on the earlier part of the curve.

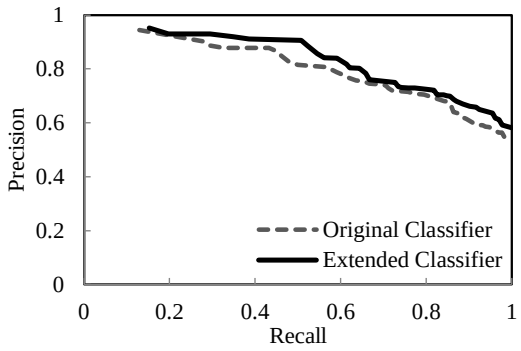

Figure 3: The new features improve precision/recall performance slightly on general Web text.

Finally, we evaluate our new confidence function on a set of extractions from random Web sentences. Figure 3 compares performance of the original confidence metric with the extended metric. The graph reaches recall 1.0, since we are using the total number of correct extractions to compute pseudo-recall. The new confidence function gives a small lift to the entire recall-precision curve. Overall, however, the two curves are quite similar, which is to be expected, since the complex constructions of nested and hypotheticals aren't as common in Web text.

In summary, we find the additional features to be particularly helpful for sentences in biology text. We expect that this new confidence function will be robust and aid other genres involving complex sentence constructions.

5. QUESTION GENERATION

In this section, we demonstrate the utility of Open IE relations for an end task. We use ReVerb tuples to automatically generate questions from the biology text book. Automatically generated questions are valuable for students in both learning and in evaluating their progress [21]. At a high-level, existing techniques use a two-step process to generate questions from Web or newswire texts [10, 11]: 1) Over-generate questions through rule-based transformations of candidate answer phrases that are extracted using entity tagging or semantic role labeling. 2) Re-rank the over-generated questions using contextual information to prune bad questions.

Approach - Different from prior work, we use relational tuples obtained from Open IE for generating questions. This enables application to new domains without requiring domain specific vocabulary of relations and also provides a simple mechanism for generating questions. A relational tuple (Arg1, Relation, Arg2) can be transformed into questions by treating one or more elements in the relational tuple as a missing value. We consider two such transformations: 1) What [exhibits] *Relation* [with] Arg2? and 2) Arg1 [exhibits] *Relation* [with] what?.[5]

For example, the tuple (Mitosis, is usually followed by, cytokinesis), will be transformed into two questions: 1) *Mi-*

[5]We leave investigation of "*Arg1* exhibits what with *Arg2*" and other possibilities for future work.

tosis is usually followed by what?, and 2) *What is usually followed by cytokinesis?*.

Evaluation - We generated questions from two chapters of our textbook. ReVerb extracted 686 relational tuples from these chapters. After filtering out tuples whose arguments had pronouns, references or questions words, we generated a total of 1206 questions, which were then annotated by four biology grad students. A question was labeled *useful* if it was a meaningful question, and has educational value for students who are learning from the biology text book. We then aggregated the annotations to label a question useful if at least two annotators agreed on it.

Out of the 1206 over-generated questions, the annotators judged 399 questions as *useful* questions according to our criteria. In other words, about 1 in 3 questions were found to be useful. Because questions generated from bad extractions are unlikely to be meaningful, we use the confidence score of the underlying extraction as a re-ranking measure. After re-ranking, we find that more than half of the top 100 questions are useful and nearly 60% of the top 50 are useful.

Error Analysis - An analysis shows that there are three primary sources of errors: 1) *Non-assertive relations* – Those with relation phrases that have modals such as 'might be coated with' or 'can be a part of' lead to bad questions that are not specific enough. These can be handled through simple modal detecting features. 2) *Non-functional relations* – Relations such as 'is one aspect of' and 'is part of' can take on many distinct arguments leading to vague questions that can have many possible answers, *e.g.*the question "What becomes part of a cell membrane?" has many answers including 'proteins' and 'vesicle'. Determining the functionality of relations is a harder problem, though recent work in the context of Web text has shown promise [15]. 3) *Overly General relations* – Finally, overly general relations or arguments also produce questions that are not clear enough to be answered, *e.g.*, "What is the cell?" and "What are models?" Determining generality or appropriate level of specificity is even more difficult and may require domain-dependent information.

In summary, we investigated the use of out-of-the-box Open IE for the task of question generation. To our knowledge, we are the first to a) investigate the use of Open IE relational tuples for question generation, b) apply it to a scientific domain, and c) evaluate it based on the educational value of the questions in the domain. Our first results are modest, though encouraging. Our analysis exposes important features that can improve question generation.

6. RELATED WORK

There has been a steady trend to reduce the human effort needed in IE. The MUC conferences in the 1990's provided 1,300 annotated training documents and allowed unlimited knowledge engineering. The recent Machine Reading (MR) program [26] provided on the order of 100 training documents with as few as 20 training instances per relation and a goal of eventually reducing knowledge engineering to a single week. One participant in the MR program explored handling a surprise pharmaceutical domain with only 50 hours of manual effort [8]. A combination of handwritten rules and learned patterns gave them recall 0.49 at precision 0.52 on a question answering task. Our current work is pushing this trend to its logical conclusion with out-of-the-box extraction.

While Open IE extracts relational tuples that are deemed useful by our domain experts, the relations and argument values are text phrases rather than concepts in a formal ontology. Preliminary work has been done to automatically map Open IE tuples to a domain ontology [25]. The system achieved precision above 0.90 with recall over 0.30 for a majority of relations in an NFL football domain from a few dozen training instances.

There has been extensive work in medical and biomedical information extraction, but this body work has relied heavily on manually-crafted knowledge and hand-tagged sentences. A series of BioNLP conferences has a shared tasks of extracting a set of relations concerning the behavior of biomolecules. [13, 17, 14]. Other shared tasks have included of extracting information about medications from clinical reports [27, 30], and extracting a set of relations between medical problems and tests or treatments from clinical reports [28]. In each of these shared tasks, the systems are provided with biomedical ontologies such as GENIA or UMLS and exhaustively tagged corpora. With their focus on predefined concepts and relations, all of these systems would have negligible recall on the biology textbook sentences.

7. CONCLUSIONS AND FUTURE WORK

This paper assessed the state of the art in applying information extraction out-of-the-box to new domains *without* any manual effort or domain-specific engineering. We utilized the state-of-the-art Open IE system ReVerb because it is readily applicable in situations where the relations of interest are not known in advance. We evaluated ReVerb's out-of-the-box performance on two bio-medical corpora – MedLine abstracts and a biology textbook. At precision 0.9, ReVerb achieved recall of 0.29 for MedLine and recall 0.14 for the biology textbook. At precision 0.7, recall jumped to 0.57 for the biology textbook (Figure 1).

Error analysis exposed common sentence constructions that were not handled well by ReVerb, which suggested new features to the confidence classifier to identify conditionally true and n-ary relations. This improvement to ReVerb doubled recall on the biology textbook at precision 0.9, and also yielded small gains on our Web text and MedLine corpora demonstrating that the improvements were not overfitting the biology domain. Finally, we investigated the first use of Open IE for the end-task of question generation, and find that our preliminary results are encouraging.

While this paper focused on detecting potential errors due to n-ary and conditionally true relations, in future work, we plan to extend the extraction process to output such relations accurately.

Acknowledgements

We are grateful to Vulcan Inc., who funded this work as part of Project Halo and provided access to the Biology textbook used this work.

8. REFERENCES

[1] ARPA. *Proc. 3rd Message Understanding Conf.* Morgan Kaufmann, 1991.

[2] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In *Procs. of IJCAI*, 2007.

[3] J. Berant, I. Dagan, and J. Goldberger. Global learning of typed entailment rules. In *Proceedings of ACL 2011*, 2011.

[4] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the Web. *Communications of the ACM*, 51(3):68–74, 2008.

[5] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[6] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. Open information extraction: the second generation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '11)*, 2011.

[7] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of EMNLP*, 2011.

[8] M. Freedman, L. Ramshaw, E. Boschee, R. Gabbard, G. Kratkiewicz, N. Ward, and R. Weischedel. Extreme extraction - machine reading in a week. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, 2011.

[9] D. Gunning, V. K. Chaudhri, P. Clark, K. Barker, S.-Y. Chaw, M. Greaves, B. Grosof, A. Leung, D. McDonald, S. Mishra, J. Pacheco, B. Porter, A. Spaulding, D. Tecuci, and J. Tien. Project HALO update – progress toward Digital Aristotle. *AI Magazine*, 31(3), 2010.

[10] C. Gütl, K. Lankmayr, J. Weinhofer, and M. Höfler. Enhanced Automatic Question Creator–EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, 9(1):23–38, 2011.

[11] M. Heilman. *Automatic Factual Question Generation from Text*. PhD thesis, Carnegie Mellon University, 2011.

[12] R. Hoffman, C. Zhang, and D. S. Weld. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, 2010.

[13] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. Overview of BioNLP-09 shared task on event extraction. In *Proceedings of the 2009 BioNLP Workshop*, 2009.

[14] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii. Overview of BioNLP shared task 2011. In *Proceedings of the 2011 BioNLP Workshop*, 2011.

[15] T. Lin, Mausam, and O. Etzioni. Identifying functional relations in Web text. In *Proceedings of EMNLP 2010*, pages 1266–1276, 2010.

[16] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *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 (ACL-IJCNLP '09)*, 2009.

[17] H. Poon and L. Vandervende. Joint inference for knowledge extraction from biomedical literature. In *Proceedings of ACL 2010*, 2010.

[18] J. B. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and R. B. Jackson. *Campbell Biology (9th Edition)*. Benjamin Cummings, 2010.

[19] E. Riloff. Automatically constructing extraction patterns from untagged text. In *Procs. of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pages 1044–1049, 1996.

[20] A. Ritter, Mausam, and O. Etzioni. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of ACL 2010*, 2010.

[21] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan. Overview of the first question generation shared task evaluation challenge. In *Proceedings of QG2010: The Third Workshop on Question Generation*, page 45, 2010.

[22] S. Schoenmackers, J. Davis, O. Etzioni, and D. S. Weld. Learning first-order Horn clauses from Web text. In *Proceedings of EMNLP 2010*, 2010.

[23] Y. Shinyama and S. Sekine. Preemptive information extraction using unrestricted relation discovery. In *Procs. of HLT/NAACL*, 2006.

[24] S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. CRYSTAL: Inducing a conceptual dictionary. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 1314–21, 1995.

[25] S. Soderland, B. Roof, B. Qin, S. Xu, Mausam, and O. Etzioni. Adapting open information extraction to domain-specific relations. *AI Magazine*, 31(3):93–102, 2010.

[26] S. Strassel, D. Adams, H. Goldberg, J. Herr, R. Keesing, D. Oblinger, H. Simpson, R. Schrag, and J. Wright. The DARPA Machine Reading Program - encouraging linguistic and reasoning research with a series of reading tasks. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

[27] O. Uzuner, I. Solti, and E. Cadag. Extracting medication information from clinical text. *Journal of American Medical Informatics Association*, 2010.

[28] O. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of American Medical Informatics Association*, 2011.

[29] F. Wu and D. S. Weld. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, 2010.

[30] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. Medex: a medical information extraction system for clinical narratives. *Journal of American Medical Informatics Association*, 17:19–24, 2010.

[31] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen. StatSnowball: a statistical approach to extracting entity relationships. In *WWW '09: Proceedings of the 18th international conference on World Wide Web*, pages 101–110, New York, NY, USA, 2009. ACM.