REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE		3. DATES COVERED *(From - To)*
03-10-2011	Final Technical Report		01-01-2008 - 12-31-2010

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Open Information Extraction	
	5b. GRANT NUMBER
	N00014-08-1-0431
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Oren Etzioni	
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	B. PERFORMING ORGANIZATION REPORT NUMBER
University of Washington Office of Sponsored Programs 4333 Brooklyn Ave NE Seattle, WA 98195-9472	UW budget 61-6549 UW eGC1 A34931

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Office of Naval Research 875 N Randolph Street Arlington, VA 22203-1995	
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Public distribution/availability

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Traditionally, Information Extraction (IE) has focused on satisfying precise, narrow, pre-specified requests from small homogeneous corpora (e.g., extract the location and time of seminars from a set of announcements). Shifting to a new domain requires the user to name the target relations and to manually create new extraction rules or hand-tag new training examples. This manual labor scales linearly with the number of target relations. This proposal introduces Open IE, a new extraction paradigm where the system makes a single data-driven pass over its corpus and extracts a large set of relational tuples without requiring any human input. The proposal also introduces TextRunner, a fully implemented, highly scalable Open IE system where the tuples are assigned a probability and indexed to support efficient extraction and exploration via user queries. Open IE is a very recent research breakthrough funded, in part, by our previous ONR grant on "Semantic Tractability on the World Wide Web". Here, we propose to study its efficacy and extend it in some important ways.

15. SUBJECT TERMS

Information Extraction, Unsupervised Learning, Self-Supervised Learning, Machine Reading, Web-Scale Extraction, Knowledge Acquisition, Query Optimization, Machine Learning, Natural Language Processing, Question Answering, Opinion Mining, Fact Checking, Inference.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Oren Etzioni
UU	UU	UU	UU	8	19b. TELEPHONE NUMBER *(Include area code)* 206-685-3035

Final Report: N00014-08-1-0431
Report Period: January 1, 2008 – December 31, 2010
Open Information Extraction

Program Officer:	Behzad Kamgar-Parsi
Sponsoring Organization:	Office of Naval Research
Address:	875 N Randolph Street, Suite 1425
City, State, Zip:	Arlington, VA 22203-1995

Principal Investigator:	Oren Etzioni
Performing Organization:	University of Washington
Address 1:	Turing Center
Address 2:	Computer Science & Engineering, Box 352350
City, State, Zip:	Seattle, WA 98195-2350
Phone Number:	206-685-3035
Fax Number:	206-543-2969
Email:	etzioni@cs.washington.edu
Website:	http://www.es.washington.edu/homes/etzioni/

Technical Section

The goal of this research was to extract knowledge from text collections as large and diverse as the Web without any human input. We produced TextRunner, an *Open Information Extraction* system that mines massive, heterogeneous text corpora to extract relational tuples without any relation-specific input or training data.

- We introduced simple syntactic and lexical constraints on how binary relationships are expressed via verbs in English sentences. The syntactic constraint is captured by a compact regular expression over parts of speech, and the lexical constraint is enforced by statistics computed over the Google N-gram corpus.

- We implemented the constraints in the OCCAM extractor, which achieves substantially improved precision/recall over state-of-the-art open extractors such as TEXTRUNNER and WOE. OCCAM has more than twice the area under the precision-recall curve compared with TEXTRUNNER, 47% more than WOEpos, and 17% more than WOEparse. OCCAM is also 30x faster than WOEparse on average.

- We developed a contradiction detection system called AuContraire, which can find contradictions between various facts present in the web text. It applies probabilistic inference over information of meronymy, functionality of relations, and ambiguity in entities to distinguish between apparent contradictions and true contradictions.

- We developed Grounder, an entity resolution system, which maps surface forms of named entities into known entities in the Wikipedia taxonomy. The new approach is based on probabilistic techniques that combine evidence from the prior popularity of entities as well as the similarity between the Wikipedia page and the current webpage.

- We developed a fact re-ranker for TextRunner that, given a set of facts for a query, computes a best order in which they should be presented to the user. This system is based upon classifiers that classify whether a fact is basic to the query, whether the fact is surprising and unexpected, etc. The basic facts and surprising facts are ranked higher. These classifiers generalize from a limited set of training data. Additionally, our system is able to personalize the search results based on data provided by a user in the past queries.

- We developed a novel hypernym extractor that combines lexico-syntactic patterns with probabilistic techniques such as Hidden Markov models to infer whether an entity pair is in hypernym-hyponym relationship.

- We are currently investigating increasing the precision and recall of TextRunner by incorporating additional linguistic resources and information. We are implementing a separate classifier for each linguistic construct such as appositives, relative clauses etc. We hope that a more precise and comprehensive TextRunner system will immensely benefit all research that uses TextRunner extractions as input.

- We are currently building a comprehensive repository of selectional preferences for arguments of each predicate based on statistical analysis over TextRunner extractions. Our preliminary results are very promising and we expect to release the data for thousands of predicates in the near future.

- We are investigating the next generation information extractor that will automatically build an expectation for possible future extractions as it reads text. For example, based on the current extraction it may add a template extraction in the database for all objects of the same type. This repository of templates will help guide the later extractions as more complex text is read.

- AuContraire discovered a surprising characterization of contradictions on the Web: of the seeming contradictions (extractions of a functional relation whose argument values disagree), only 1.2% are actual contradictions, from a set of TextRunner extractions from 117 million Web pages. The false contradictions have argument values that are compatible due to synonymy or meronymy (e.g. 'Vienna' does not contradict 'Austria'). Ambiguous argument values that refer to different real-world entities also produce false contradictions. Despite the badly skewed data, AuContraire found true contradictions with precision 1.0 at recall 0.15 and with precision 0.48 at recall 0.29.

- AuContraire learned functionality of predicates and ambiguity of arguments in alternating EM-like iterations. It achieved precision 0.67 at recall 0.55 for functionality and precision 0.87 at recall 0.34 for ambiguity.

- Grounder demonstrated the importance of prior probabilities in mapping references in context to Wikipedia articles. Cosine similarity between a document and the Wikipedia article gave precision 0.67 at recall 0.27, while a prior that ignores context gave precision 1.0 at recall 0.31. Combining both sources of knowledge gave results superior to either alone, achieving precision 0.91 at recall 0.62.

- Our HypernymFinder found at least one correct hypernym for proper nouns with precision 0.90 at recall 0.32 (as compared with WordNet that covered only 17% of the proper nouns in our test set). For common nouns HypernymFinder had precision 0.90 at recall 0.67. An HMM-based classifier handles instances not covered by lexico-syntactic patterns, increasing recall by 0.06 for proper nouns and by 0.02 for common nouns.

- Our fact re-ranker evaluated several definitions of interestingness of a fact. We found that three definitions, basic facts, specific facts and distinguishing facts, can make a fact interesting. Often these span different kinds of facts. Our re-ranker was able to increase the number of interesting facts in the first thirty results of the query from 42% to 64% resulting in a better user experience for the users.

- We reimplemented TextRunner's tuple extractor using a self-supervised Conditional Random Field (CRF). TextRunner learns a relation-independent extractor by automatically generating positive and negative training examples from parse trees and a small set of relation-independent heuristics. Where the previous TextRunner was limited to binary tuples of the form (arg1, pred, arg2), the new implementation finds tuples with an arbitrary number of arguments.

- We evaluated TextRunner's open extraction model relative to the traditional extraction paradigm in which a relation is specified in an advance, along with hand-labeled training data per relation.

- We built the Holmes system on top of TextRunner, which is able to infer new facts not seen on any page in the corpus. It does this by combining facts from multiple web pages using a small set of rules. Furthermore we demonstrated that relations extracted from the web have the property of being Approximately Pseudo-functional (most entities appear with only a small number of other entities), and this property allows Holmes's inference to scale linearly with the size of the input corpus. For some example queries, we demonstrated that Holmes doubled recall over the baseline TextRunner system, and can do so in only a few CPU minutes.

- We developed Alice, one of the first learning agents whose goal is to automatically discover a domain theory – a collection of concepts, facts and generalizations for a given topic -- directly from Web text. Alice uses relational tuples extracted by TextRunner to learn new concepts and build relationships between concepts in a hierarchy.

- We demonstrated that the new implementation of TextRunner can extract a variety of relations with precision 88.3% and recall 45.2%, while the previous implementation had precision 86.6% at recall 23.2%. This is gives an F1 measure 63.4% higher than the previous implementation.

- We found that without any relation-specific input, TextRunner obtains the same precision with lower recall as a traditional supervised extractor trained using hundreds and sometimes thousands, of labeled examples per relation.

- We are currently observing that relational tuples located by TextRunner can be used to bootstrap training of individual relations. TextRunner automatically provides several orders of magnitude more training data without the cost of hand-tagging, yielding substantial gains in F1 on a per-relation basis.

Awards and Honors

Stefan Schoenmackers' Ph.D. dissertation was accepted in Winter 2011, but a copy is not yet available.

"Unsupervised named-entity extraction from the Web: An experimental study" is the most cited *Artificial Intelligence* article in the last 5 years.

Ph.D. alum Doug Downey selected as a Microsoft Research Faculty Fellow in 2010.

Oren Etzioni awarded a Washington Research Foundation Endowed Entrepreneurship Professorship in Computer Science & Engineering in 2009.

Michael Cafarella received his Ph.D. degree in Summer 2009; his dissertation is included in the publications section below.

Michele Banko received her Ph.D. degree in Spring 2009; her dissertation is included in the publications section below.

Doug Downey received his Ph.D. degree in Autumn 2008; his dissertation is included in the publications section below.

Thomas Lin and Alan Ritter were each awarded 3 year National Defense Science and Engineering Graduate (NDSEG) Fellowships in 2008.

Anthony Fader was awarded a 3 year National Science Foundation Graduate Fellowship in 2007.

Patent Filings or Patent Awards

There were no patent filings or patent awards resulting from this grant. However a patent award was received during the performance period of this grant. That patent award, identified below, resulted from a patent filing that was submitted during our previous grant, N00014-05-1-0185.

US patent 7,877,343
Title: Open Information Extraction From the Web
Awarded: January 25, 2011

Publications

Publications produced during the report period are listed below; online copies of these publications can be found at http://turing.cs.washington.edu/publications.htm.

1. Thomas Lin, Mausam, and Oren Etzioni, "Commonsense from the Web: Relation Properties," in Proceedings of the 2010 AAAI Fall Symposium on Commonsense Knowledge.

2. Thomas Lin, Mausam, and Oren Etzioni, "Identifying Functional Relations in Web Text," in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

3. Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, and Jesse Davis, "Learning First-Order Horn Clauses from Web Text," in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

4. Doug Downey, Oren Etzioni, and Stephen Soderland, "Analysis of a Probabilistic Model of Redundancy in Unsupervised Information Extraction," *Artificial Intelligence*, 174(11): 726-748, 2010.

5. Anthony Fader, Stephen Soderland, and Oren Etzioni, "Extracting Sequences from the Web," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

6. Alan Ritter, Mausam, and Oren Etzioni, "A Latent Dirichlet Allocation method for Selectional Preferences," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

7. Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni, "Semantie Role Labeling for Open Information Extraetion," in Proeeedings of the Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR) at NAACL 2010.

8. Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Mausam, Alan Ritter, Stefan Sehoenmaekers, Stephen Soderland, Dan Weld, Fei Wu, and Congle Zhang, "Maehine Reading at the University of Washington," in Proeeedings of the Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR) at NAACL 2010.

9. Thomas Lin, Oren Etzioni, and James Fogarty, "Identifying Interesting Assertions from the Web," in Proeeedings of the 18th Conferenee on Information and Knowledge Management, 2009.

10. Alan Ritter, Stephen Soderland, and Oren Etzioni, "What Is This, Anyway: Automatie Hypernym Diseovery," in Proeeedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read.

11. Alexander Yates and Oren Etzioni, "Unsupervised Methods for Determining Objeet and Relation Synonyms on the Web," *Journal of Artificial Intelligence Research*, 34: 255-296, 2009.

12. Miehael Cafarella, "Extraeting and Managing Struetured Web Data," *University of Washington Ph.D. Dissertation*, 2009.

13. Miehele Banko, "Open Information Extraetion for the Web," *University of Washington Ph.D. Dissertation*, 2009.

14. Oren Etzioni, Miehele Banko, Stephen Soderland, and Daniel S. Weld, "Open Information Extraetion from the Web," *Communications of the ACM*, 51(12): 68-74, 2008.

15. Doug Downey and Oren Etzioni, "Look Ma, No Hands: Analyzing the Monotonie Feature Abstraetion for Text Classifieation," in Proeeedings of the 22nd Annual Conferenee on Neural Information Proeessing Systems, 2008.

16. Bhushan Mandhani and Stephen Soderland, "Exploiting Hyponymy in Extraeting Relations and Enhaneing Ontologies," in Proeeedings of the IEEE/WIC/ACM Conferenee on Web Intelligenee Workshop on Natural Language Proeessing and Ontology Engineering, 2008.

17. Doug Downey, "Redundaney in Web-seale Information Extraetion: Probabilistie Model and Experimental Results," *University of Washington Ph.D. Dissertation*, 2008.

18. Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni, "It's a Contradiction – no, it's not: A Case Study using Functional Relations," in Proceedings of the Conference on Empirical Methods on Natural Language Processing, 2008.

19. Stefan Schoenmackers, Oren Etzioni, and Daniel Weld, "Scaling Textual Inference to the Web," in Proceedings of the Conference on Empirical Methods on Natural Language Processing, 2008.

20. Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni, "It's a Contradiction – no, it's not: A Case Study using Functional Relations," to appear in Proceedings of the Conference on Empirical Methods on Natural Language Processing, 2008.

21. Stefan Schoenmackers, Oren Etzioni, and Daniel Weld, "Scaling Textual Inference to the Web," to appear in Proceedings of the Conference on Empirical Methods on Natural Language Processing, 2008.

22. Michele Banko and Oren Etzioni, "The Tradeoffs Between Open and Traditional Relation Extraction," in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008.

Presentations

1. Invited Talk (International Joint Conference on Artificial Intelligence, Barcelona, Spain), "Open Information Extraction at Web Scale." July 2011.

2. Invited Talk (IBM Information Retrieval Seminar, Haifa, Israel), "Open Information Extraction at Web Scale." December 2010.

3. Invited Talk (Colloquium on Digital Humanities and Computer Science, University of Chicago), "Machine Reading of the World Wide Web." November 2008.

4. Invited Talk (conference on Empirical Methods in Natural Language Processing, Honolulu, HI), "We KnowItAll: lessons from a quarter century of Web extraction research." October 2008.

5. Invited Talk (Web Search and Data Mining, Stanford, CA), "Machine Reading at Web Scale." February 2008.

People Supported (Faculty, Students, Technical Staff)

1. Oren Etzioni, Faculty, Principal Investigator
2. Michele Banko, Graduate Student
3. Bo Qin, Graduate Student
4. Mausam, Research Faculty
5. Stephen Soderland, Research Scientist
6. Thomas Lin, Graduate Student
7. Alan Ritter, Graduate Student
8. Yoav Artzi, Graduate Student

Project and Related Websites

http://turing.cs.washington.edu/
http://www.cs.washington.edu/research/knowitall/
http://ai.cs.washington.edu/projects/open-information-extraction
http://www.cs.washington.edu/research/textrunner/
http://reverb.cs.washington.edu/
http://turinge.cs.washington.edu:1234/lda_sp_demo_v3/lda_sp/relations/
http://abstract.cs.washington.edu/~tlin/leibniz/