# TextRunner: Open Information Extraction on the Web

**Alexander Yates** **Michele Banko** **Matthew Broadhead**
**Michael Cafarella** **Oren Etzioni** **Stephen Soderland**
University of Washington
Computer Science and Engineering
Box 352350
Seattle, WA 98195-2350
{ayates,banko,hastur,mjc,etzioni,soderlan}@cs.washington.edu

## 1 Introduction

Traditional information extraction systems have focused on satisfying precise, narrow, pre-specified requests from small, homogeneous corpora. In contrast, the TextRunner system demonstrates a new kind of information extraction, called Open Information Extraction (OIE), in which the system makes a single, data-driven pass over the entire corpus and extracts a large set of relational tuples, without requiring *any* human input. (Banko et al., 2007) TextRunner is a fully-implemented, highly scalable example of OIE. TextRunner's extractions are indexed, allowing a fast query mechanism.

Our first public demonstration of the TextRunner system shows the results of performing OIE on a set of 117 million web pages. It demonstrates the power of TextRunner in terms of the raw number of facts it has extracted, as well as its precision using our novel assessment mechanism. And it shows the ability to automatically determine synonymous relations and objects using large sets of extractions. We have built a fast user interface for querying the results.

## 2 Previous Work

The bulk of previous information extraction work uses hand-labeled data or hand-crafted patterns to enable relation-specific extraction (e.g., (Culotta et al., 2006)). OIE seeks to avoid these requirements for human input.

Shinyama and Sekine (Shinyama and Sekine, 2006) describe an approach to "unrestricted relation discovery" that does away with many of the requirements for human input. However, it requires clustering of the documents used for extraction, and thus scales in quadratic time in the number of documents. It does not scale to the size of the Web.

For a full discussion of previous work, please see (Banko et al., 2007), or see (Yates and Etzioni, 2007) for work relating to synonym resolution.

## 3 Open IE in TextRunner

OIE presents significant new challenges for information extraction systems, including

**Automation** of relation extraction, which in traditional information extraction uses hand-labeled inputs.

**Corpus Heterogeneity** on the Web, which makes tools like parsers and named-entity taggers less accurate because the corpus is different from the data used to train the tools.

**Scalability** and efficiency of the system. Open IE systems are effectively restricted to a single, fast pass over the data so that they can scale to huge document collections.

In response to these challenges, TextRunner includes several novel components, which we now summarize (see (Banko et al., 2007) for details).

**1. Single Pass Extractor**

The TextRunner extractor makes a single pass over all documents, tagging sentences with part-of-speech tags and noun-phrase chunks as it goes. For each pair of noun phrases that are not too far apart, and subject to several other constraints, it applies a classifier described below to determine whether or not to extract a relationship. If the classifier



*NAACL HLT Demonstration Program*, pages 25–26,
Rochester, New York, USA, April 2007. 

deems the relationship trustworthy, a tuple of the form $t = (e_i, r_j, e_k)$ is extracted, where $e_i, e_k$ are entities and $r_j$ is the relation between them. For example, TEXTRUNNER might extract the tuple *(Edison, invented, light bulbs)*. On our test corpus (a 9 million document subset of our full corpus), it took less than 68 CPU hours to process the 133 million sentences. The process is easily parallelized, and took only 4 hours to run on our cluster.

**2. Self-Supervised Classifier**

While full parsing is too expensive to apply to the Web, we use a parser to generate training examples for extraction. Using several heuristic constraints, we automatically label a set of parsed sentences as trustworthy or untrustworthy extractions (positive and negative examples, respectively). The classifier is trained on these examples, using features such as the part of speech tags on the words in the relation. The classifier is then able to decide whether a sequence of POS-tagged words is a correct extraction with high accuracy.

**3. Synonym Resolution**

Because TEXTRUNNER has no pre-defined relations, it may extract many different strings representing the same relation. Also, as with all information extraction systems, it can extract multiple names for the same object. The RESOLVER system performs an unsupervised clustering of TEXTRUNNER's extractions to create sets of synonymous entities and relations. RESOLVER uses a novel, unsupervised probabilistic model to determine the probability that any pair of strings is co-referential, given the tuples that each string was extracted with. (Yates and Etzioni, 2007)

**4. Query Interface**

TEXTRUNNER builds an inverted index of the extracted tuples, and spreads it across a cluster of machines. This architecture supports fast, interactive, and powerful relational queries. Users may enter words in a relation or entity, and TEXTRUNNER quickly returns the entire set of extractions matching the query. For example, a query for "Newton" will return tuples like *(Newton, invented, calculus)*. Users may opt to query for all tuples matching synonyms of the keyword input, and may also opt to merge all tuples returned by a query into sets of tuples that are deemed synonymous.

## 4 Experimental Results

On our test corpus of 9 million Web documents, TEXTRUNNER extracted 7.8 million well-formed tuples. On a randomly selected subset of 400 tuples, 80.4% were deemed correct by human reviewers.

We performed a head-to-head comparison with a state-of-the-art traditional information extraction system, called KNOWITALL. (Etzioni et al., 2005) On a set of ten high-frequency relations, TEXTRUNNER found nearly as many correct extractions as KNOWITALL (11,631 to 11,476), while reducing the error rate of KNOWITALL by 33% (18% to 12%).

## Acknowledgements

This research was supported in part by NSF grants IIS-0535284 and IIS-0312988, DARPA contract NBCHD030010, ONR grant N00014-05-1-0185 as well as gifts from Google, and carried out at the University of Washington's Turing Center.

## References

M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open Information Extraction from the Web. In *IJCAI*.

A. Culotta, A. McCallum, and J. Betz. 2006. Integrating Probabilistic Extraction Models and Relational Data Mining to Discover Relations and Patterns in Text. In *HLT-NAACL*.

O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91–134.

Y. Shinyama and S. Sekine. 2006. Preemptive Information Extraction Using Unrestricted Relation Discovery. In *HLT-NAACL*.

A. Yates and O. Etzioni. 2007. Unsupervised Resolution of Objects and Relations on the Web. In *NAACL-HLT*.