

# Marve: Measurement Context Extraction from Text

Discovering Opportunities for NASA Jet Propulsion Laboratory’s Proposed HypsIRI Mission

Kyle Hundman<sup>1</sup>, Chris A. Mattmann<sup>1,2</sup>  
{kyle.a.hundman,chris.a.mattmann}@jpl.nasa.gov

<sup>1</sup>Jet Propulsion Laboratory  
California Institute of Technology  
Pasadena, CA 91109 USA

<sup>2</sup>Computer Science Department  
University of Southern California  
Los Angeles, CA 90089 USA

## ABSTRACT

We propose Marve, a system for extracting measurement values, units, and related words from natural language text. Marve uses conditional random fields (CRF) to identify measurement values and units, followed by a rule-based system to find related entities, descriptors and modifiers within a sentence. Sentence tokens are represented by an undirected graphical model, and rules are based on part-of-speech and word dependency patterns connecting values and units to contextual words. Marve is unique in its focus on measurement context and early experimentation demonstrates Marve’s ability to generate high-precision extractions with strong recall. We also discuss Marve’s role in justifying NASA JPL’s proposed HypsIRI mission, a hyperspectral infrared imaging satellite that will study the world’s ecosystems. In general, our work with HypsIRI demonstrates the value of semantic measurement extractions in characterizing quantitative discussion contained in large corpuses of natural language text. These extractions accelerate broad-cross cutting literature surveys and expose researchers and scientists new algorithmic approaches and experimental nuances. They also facilitate identification of scientific opportunities enabled by HypsIRI leading to more informed scientific investment and research.

## CCS CONCEPTS

•**Computing methodologies** → **Information extraction**; Maximum likelihood modeling; •**Applied computing** → **Environmental sciences**; Bioinformatics; •**Information systems** → *Content analysis and feature selection*;

## KEYWORDS

Natural Language Processing, Knowledge Discovery, Information Retrieval, Measurement Extraction, Natural Language Understanding, Remote Sensing, Earth Science, Machine Learning

### ACM Reference format:

Kyle Hundman<sup>1</sup>, Chris A. Mattmann<sup>1,2</sup>. 2017. Marve: Measurement Context Extraction from Text. In *Proceedings of 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada, August 2017 (KDD 2017)*, 9 pages.  
DOI: 10.475/123.4

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD 2017, Halifax, Nova Scotia, Canada

© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06.

DOI: 10.475/123.4

## 1 INTRODUCTION

Much of the world’s scientific information is easily accessible from our fingertips. For example, a search for “remote sensing” on Thomson Reuter’s Web of Science yields nearly 14,000 journal articles from 2014–2016<sup>1</sup>. However, the careful analysis and understanding of that scientific information is not in and of itself, easy. An individual scientist may spend many hours reading, and re-reading a single article just to comprehend what it is trying to say. In the prior example about remote sensing the aforementioned corpus of articles returned is simply too large for a human, or even a small set of them, to read each and every article and to synthesize the resultant information into knowledge from them. Fortunately, continual advances in search and natural language processing (NLP) have greatly enhanced our ability to automatically characterize and sift through large-scale unstructured data. Neural network approaches have vastly improved essential NLP tasks such as part-of-speech (POS) tagging and dependency parsing. For example, Stanford CoreNLP’s dependency parser achieved 91.7% [8] accuracy on the Penn Treebank dataset and Google’s Parsey McParseface attained an impressive 94.4% on sentences from various news sources [3].

While NLP provides a foundational framework for processing text, scientifically understanding that text requires more. One of the core elements of science is *measurement* which involves quantification (in some units e.g., nanometers; microns, etc.); possibly a min/max range, understanding of the measured value, etc. Measurements abound in scientific literature. Consider the sentence, “The unexpected drop in stratospheric water vapor slowed the rate of increase in surface temperature in the subsequent decade by 25%.” Identifying “25%” as the “rate of increase in surface temperature” is difficult without a system that considers the underlying phrasal structure of the sentence. Proper measurement extraction and labeling enables the creation of unique knowledge bases and opens exciting possibilities for modeling and visualization techniques that rely on organized and uniform numerical data. In the 14,000 journal articles already described, automatically discerning the scientific measurements; their values, and min/max range can allow for quick summarization and scientific understanding of a large corpus of literature and/or news articles. In addition, it can allow for validation and comparison of automated textual science understanding with that of remotely sensed measurements, providing additional context, and even scientific corroboration of phenomena.

We describe a framework, Marve, that fuses existing techniques in NLP and text processing to extract context around measurements

<sup>1</sup><https://apps.webofknowledge.com>

in natural language. Using rules primarily based on word dependencies and POS tags, Marve exploits a limited set of English language patterns used to discuss measurements and the objects or concepts they quantify. Traditionally, the cost of manual curation and the ambiguity of unaccompanied measurements have limited the collection and application of semantic measurement data. For instance, measurement context enables improved identification of measurement types. From a scientific perspective, Marve accelerates exploration of literature and promote cross-pollination of ideas and approaches across domains.

## 2 MOTIVATION

Marve originated from a NASA Advanced Concepts project that has provided data-driven support for NASA JPL’s proposed Hyperspectral Infrared Imager (HyspIRI) mission. HyspIRI is a hyperspectral infrared imager that will monitor a variety of ecological and geological features at a wide range of wavelengths. The planned HyspIRI instrumentation has unique technical capabilities such as high spatial resolution and hyperspectral coverage that will benefit various scientific areas [11]. However, the extent and nature of these benefits are not comprehensively understood because much of this information is embedded within scientific publications spread across numerous journals. We set out to automatically identify and profile these new scientific opportunities using a corpus of approximately 2,500 recent publications and abstracts from various journals in the remote sensing domain.

Our first approach involved the use of regular expressions to extract common measurement types in the remote sensing domain. Extracted measurements like *spatial resolution*, *spectral coverage* and *revisit rate* provided a useful bookmarking of our corpus – discussions around hyperspectral wavelengths (>2400 nm) and high spatial resolutions were pointers toward potential science enabled by HyspIRI. Visualizing these extractions also revealed the scale (number of mentions) of the discussion around various wavelengths. Unfortunately, the regular expression-base approach was insufficient in several ways: precision and recall of extractions was unknown, regular expressions didn’t generalize across measurement types, and regular expressions are complex and sensitive. Most importantly, measurements were extracted in isolation. An extraction of “50 m” could be a measurement of height, swath, length, or resolution in the context of remote sensing. We attempted to look for pre-defined words in proximity to indicate context but this approach was tedious and error-laden. Ultimately, Marve resulted from our pursuit of a general, accurate, and precise tool that provided semantically-rich extractions.

## 3 RELATED WORK

### 3.1 Measurement Extraction

**3.1.1 Grobid Quantities.** The Marve stack includes Grobid Quantities, which uses linear conditional random fields (CRF) to identify measurement units and values. For both types of extraction performed by Grobid Quantities, labeled training data (ideally from the domain targeted for extractions) and subsequent model training is required. This pre-work is tolerable for measurement extraction given the alternatives. Measurement conventions and unit formats vary widely across scientific domains and the resulting proliferation

of patterns is too large and varied for unsupervised models or rule-based systems to be effective. (could put some examples here from dark matter papers). We initially tried extracting measurements using a rule-based approach built around POS tags, named-entity recognition (NER), word dependencies, and regular expressions. While this approach was fast and training-free, attempts at generalizing the system for different domains led to numerous false positive extractions. Aspects such as capitalization, abbreviations, and spacing vary across different number and unit conventions. Aras et. al [4] provide further discussion of these challenges.

Through this exploration process we determined quantified substances and related entities (i.e. context) don’t present the same challenges. They follow common language patterns that generalize well. Marve identifies words and entities that form the context around a measurement without the need for training and labeled examples. Marve can also capture related entities from a broad assortment of language patterns. Consider the following sentence: “Using Landsat-8 data, modelers achieved 82% classification accuracy for cutleaf teasal.” Grobid Quantities isn’t designed to identify “Landsat-8 data” or “cutleaf teasal” as related entities. Marve is able to capture this additional information without domain-specific labels or training - these types of phrasal patterns and clauses are common across the English language. Additionally, modifying the behavior of Marve is transparent and easily adjusted via the JSON representation of desired language patterns. While Grobid Quantities could be extended to capture more context, tuning Grobid extractions requires adding or adjusting labels and re-training the full model for a specific domain. Marve mitigates additional overhead required for context extraction.

**3.1.2 Quantalyze and GATE.** Quantalyze<sup>2</sup> is a commercial product that also performs measurement, unit, and context extraction. Evaluation of its performance can only be achieved through their online demo, but after comparing extractions from several paragraphs of text, their tool appears to achieve poor recall in both measurements captured and quantified substances captured. Additionally, Agatonovic et. al [1] employ GATE [9] to extract measurement values and units from patent documents. Their approach involves building patent-specific gazeteers and hand-written rules to generate measurement annotations using GATE. While Grobid Quantities requires labeled data, its embedded CRF model obviates rule-writing and Agatonovic et. al’s approach only captures accompanying words such as “less than” or “between” while ignoring related entities and context.

### 3.2 Open Information Extraction (OIE)

Various OIE systems approach relation extraction in similar ways to Marve. Kraken [2], CSD-IE [6] and ClausIE [10] utilize dependency patterns and POS tags to detect clauses and find their propositions. This information is then used to construct triples representing facts in a corpus, such as: (“Kelly”, “finished”, “nursing school”). Similar to the Agatonovic et. al’s GATE-based approach, certain measurements could be extracted via OIE approaches. But OIE is centered on verb-mediated prepositions and measurement context occurs in a variety of other forms such as adverbials: “The satellite captured imagery with 50 m spatial resolution.” This leads to poor OIE recall

<sup>2</sup><https://www.quantalyze.com/>

for measurements. When measurements are extracted by these systems, the output requires significant post-processing to filter extraneous OIE extractions and properly separate measurements, units, and related entities. Marve is a more directed form of these approaches, and it is most similar to those that sacrifice efficiency for improved precision and recall (e.g. OLLIE [19], Kraken, and ClausIE).

### 3.3 Relationship Extraction

Knowledge base construction (KBC) and relationship extraction have migrated from pattern matching and rule-based systems to machine learning based systems over the last decade. One of the driving factors behind this trend is that KBC systems that rely on a multitude of rules require some assurance of the precision and recall of such rules []. In practice this is difficult and tedious. As discussed in section 3.1, this is also the reason for the use of machine learning approaches to measurement value and unit extraction - too many patterns and rules result in uncertainties about precision and recall. However, Marve differs from traditional KBC systems in two primary ways. First, Marve doesn't explicitly classify types of relationships between extracted measurements and their related words and entities. Second, Marve is directed at a very specific type of extraction (measurements) that benefits many scientific information extraction scenarios. In this sense, it is complimentary to broader KBC systems. One of the most prominent KBC systems of late is DeepDive, which is designed to identify relationships between extraction types using labeling functions written by domain experts. However, generating the extractions of interest is left up to the user and writing discerning labeling functions is an gradual, iterative process. To this end, Marve automates a large part of the development of a DeepDive system by automatically extracting measurements and providing targeted measurement context to labeling functions. Section 6.2 includes further discussion around Marve and DeepDive integrations.

## 4 METHODOLOGY

### 4.1 Overview

As discussed in section 3, we decided against a custom measurement extractor and instead used Grobid Quantities to extract measurements, units, and some quantified substances. Like Marve, Grobid Quantities represents sentences with undirected graphs. Though instead of parsing language patterns, Grobid uses a probabilistic graphical CRF method that learns parameters through maximum likelihood estimation. This approach to extracting numerical values and units was more consistent in our experimentation although it adds processing overhead and requires labeled training data.

Once measurements values and units are identified using Grobid, Stanford's CoreNLP library [14] is used to perform more traditional NLP tasks such as tokenization, POS tagging, and word dependency parsing. Marve uses combinations of the output from these tasks to identify measurement types (e.g. 10 m *spatial resolution*) and related entities (e.g. *Hannibal* had around 40 elephants). These patterns originate at the measurement unit token(s) and expand outward to connected nodes (words) in the graph. If Marve finds

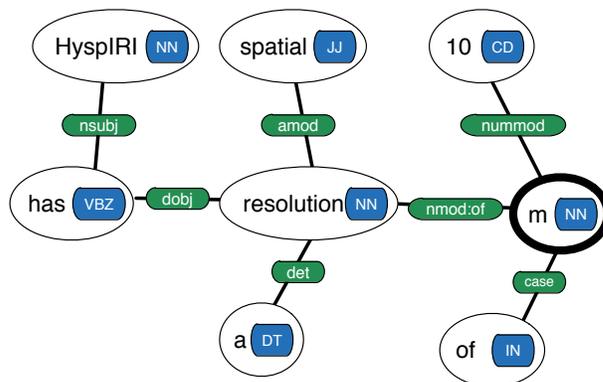


Figure 1: An example graph generated for the sentence: “HypsiRI has a spatial resolution of 10 m.” Nodes are labeled with words and POS tags and edges are labeled with word dependencies. In this case, the measurement unit “m” is the origin for the pattern matching Marve uses to identify related words and entities.

a pattern defined as valid in the hierarchical JSON framework<sup>3</sup>, it will be returned as an entity related to the measurement. To facilitate pattern evaluation, each sentence is loaded into a graph using the well-established NetworkX library<sup>4</sup> written in python.

### 4.2 Model Structure

Consider a connected, undirected graph  $G = (V, E)$  where  $V$  and  $E$  denote the sets of nodes and edges respectively, such that:

- $S = \{s_1, s_2, \dots, s_n\}$  is a set of sentences that comprise a corpus of text from which measurements are extracted.
- $T = \{t_1, t_2, \dots, t_n\}$  is a set of all tokens in  $S$ .
- $s_i = \{t \mid t \in T\}$  where each sentence  $s \in S$  is a set of  $t$  tokens.
- One graph  $G$  is constructed for each sentence  $s_i$ .
- $L = \{l_1, l_2, \dots, l_n\}$  where  $l_i$  is a label which identifies the part of speech for each  $t_i$  token in a sentence  $s_i$ .

Given these notations, the set of nodes  $V$  in each graph can be defined as  $V = \{v_1, v_2, \dots, v_n\}$  where  $v_i$  stores a token  $t_i \in t$  for a set of  $t$  tokens in a sentence  $s$ , and each token  $t_i$  is labeled with label  $l_j \in L$ . Then we define  $e_{ij}$  as an edge connecting  $(v_i, v_j)$  with a label  $d_{ij}$  representing the dependency between tokens  $(t_i, t_j)$ , where  $D$  is the set of all dependencies equal to the length of  $E$ .

### 4.3 Pattern Matching

Once a measurement is identified by Grobid Quantities, the token  $t_i$  that corresponds to the measurement unit becomes the origin for subsequent pattern evaluation (as shown by the thicker circle in Figure 1). A graph  $g_i$  is constructed for sentence  $s_i$  containing the measurement. If the dependency label  $d_{ij}$  for any edge  $e_{ij}$  originating at  $t_i$  matches the valid word dependency types defined in

<sup>3</sup>[https://github.com/khundman/marve/blob/master/marve/dependency\\_patterns.json](https://github.com/khundman/marve/blob/master/marve/dependency_patterns.json)

<sup>4</sup><https://networkx.github.io/>

**Table 1: Sample dependency pattern that would result in the extraction of “spatial resolution” as related to “10 m” from the example sentence in Figure 1. (see Table 2 for example of full extraction).**

```
{
  ``nmod``: {
    ``enhanced``: true,
    ``of``: {
      ``measurement_types``: [``space_between``, ``attached``],
      ``pos_in``: {
        ``NN``: null
      }
    }
  }
}
```

the dependency pattern JSON evaluation continues. One subtlety stems from CoreNLP’s enhanced dependencies, which provide the connecting word for certain dependency types. For example, if the conjunction “and” connects two words the enhanced dependency type returned by CoreNLP is “conj:and” rather than “conj”. When Marve encounters a dependency type that has been enhanced, it evaluates the enhanced portion separately allowing for more nuanced pattern definitions. This is represented in the JSON structure with a boolean value for the “enhanced” key (shown in Table 1).

If “enhanced” is true, the dependency label  $d_{ij}$  is split into two parts – the connecting word and dependency type. If both parts match the JSON structure, evaluation continues along that nested path.

Marve considers the format of the measurement after word dependencies are evaluated. Three primary formats are defined:

- space\_between (e.g. “10 m”)
- attached (e.g. “10m”)
- hyphenated (e.g. “10-m”)

Measurement formats are identified using the character indices of measurement value token  $t_k$  and measurement unit token  $t_i$ . If  $t_i$  and  $t_k$  are adjacent (without a space), they are “attached”. If not, a simple check for a space or hyphen is performed. Word dependency and POS patterns vary based on these formats and explicitly defining rules around them improve Marves precision.

POS tags are the next evaluation step in Marve. If the measurement format is valid given the dependency pattern and token  $t_j$  is connected to the unit token  $t_i$  via a valid dependency pattern, label  $l_j$  is evaluated in one of two ways:

- pos\_in: As long as one of the keys in the pos\_in JSON object matches part of label  $l_j$  they are valid (e.g. if one of the keys for the pos\_in nested object is “NN” and label  $l_j$  is “NNS”)
- pos\_equals: The specified POS labels must match label  $l_j$  exactly

If a matching POS key has its own keys and values in the JSON, this is where special cases are handled. Most of these involve verbs, which are often part of a clause containing a subject related to the measurement. In this case, all nodes connect to token  $t_j$  are evaluated by a separate function. Valid word dependencies are passed as parameters to this function and it executes recursively

**Table 2: Sample Marve output for the example sentence in Figure 1. The “quantity” field is populated by Grobid Quantities and the “related” field is added by Marve.**

```
{
  ``type``: ``value``,
  ``quantity``: {
    ``parsedValue``: 10,
    ``normalizedQuantity``: 10,
    ``rawValue``: ``10``,
    ``rawUnit``: {
      ``offsetStart``: 39,
      ``offsetEnd``: 40,
      ``tokenIndices``: [``8``],
      ``name``: ``m``
    },
    ``offsetEnd``: 38,
    ``offsetStart``: 36,
    ``tokenIndex``: 7,
    ``normalizedUnit``: {
      ``type``: ``length``,
      ``name``: ``m``,
      ``system``: ``SI base``
    },
    ``type``: ``length``
  },
  ``related``: [
    {
      ``rawName``: ``resolution``,
      ``connector``: ````,
      ``offsetEnd``: 32,
      ``relationForm``: ``nmod:of``,
      ``offsetStart``: 22,
      ``tokenIndex``: 5,
      ``descriptors``: [
        {
          ``rawName``: ``spatial``,
          ``tokenIndex``: ``4``
        }
      ]
    }
  ]
}
```

if it encounters connected verb tokens. If the value of a matching POS key in the JSON is null, no more evaluation is needed and the token  $t_j$  is returned as a related word.

The last step in constructing the output is finding adjectives, modifiers, or compounds connected to related nouns. This includes words like “spatial” in the example in Figure 1, where “resolution” is the related noun extracted in earlier steps. Other connected words could be subjective words like “high” in “a high spatial resolution of 10 m,” or statistical words such as “average.” These descriptive words provide important details about types, sentiment, and the statistical nature of measurements. They create opportunities for higher fidelity grouping of like measurements (e.g. “spatial resolution” versus “resolution”) and better profiling of trends and opportunities. For example, if a satellite’s 100m spatial resolution was described as “insufficient” for classifying a certain type of vegetation, this could represent an opportunity for the higher-resolution HypSPIRI mission to enable new science. Extraction of these type of words is also performed using POS labels and word dependencies, but the

**Table 3: Experiment Data**

Source	Sentences	Measurements	Sent. with Measurements
News	117	58	47
Journal	372	131	93
Total	489	189	140

origin for pattern matching is the token corresponding to a related entity rather than the measurement unit.

Marve’s rule-based architecture is transparent, flexible, and general enough to be easily modified to identify other types of relationships. Rather than providing measurement units as the origin for Marve’s pattern parsing, other types of entities could be provided. Marve only requires a token from which defined POS and dependency patterns originate.

## 5 EXPERIMENTS AND RESULTS

### 5.1 Data

Four documents were used for experimentation: two news articles from the New York Times and two scientific publications from the medical and remote sensing domains. The first New York Times article, “Dell Gets Bigger and Hewlett-Packard Gets Smaller in Separate Deals,” was selected from the Technology section and the other, “A Cleaning Start-Up Wielding Mops, Buckets and 700 Data Points” was from the business section [1]. The medical publication, “Zika Virus Associated with Microcephaly,” is from the *New England Journal of Medicine* and the remote sensing publication, “Satellite soil moisture for agricultural drought monitoring: Assessment of the SMOS derived Soil Water Deficit Index,” is from *Remote Sensing of Environment* [2]. For each document, individual sentences were manually extracted along with any measurement values, units, and related words contained within. The total amount of labeled sentences and measurements for each type of source is presented in Table 3. We avoided data sources with more informal language (e.g. social media) for two reasons: Marve will be most useful in domains with abundant quantitative discussion and Marve’s reliance on sentence structure suggests it would perform poorly on such data.

### 5.2 Setup

The labeling of related words was limited to those directly related to the measurement, most commonly connected by a verb or nominal modifiers indicating a prepositional phrase. As an example, consider the sentence in Figure 1, “HyspIRI has a spatial resolution of 10 m.” In this case, “10 m” is directly modifying “spatial resolution”, which is then possessed by “HyspIRI.” Because there is a degree of separation between “10 m” and “HyspIRI,” Marve would only include “spatial resolution” as related to “10 m.” Extracting second-order related words is easily achieved in Marve, but we focused on first-order relations to reduce manual labeling effort and simplify the evaluation.

The experiment demonstrates the precision and recall of Marve’s related word extraction independent of Grobid Quantities’ ability to accurately extract measurement values and units. Since Marve’s

**Table 4: An example of labeled evaluation data for a sentence containing a measurement.**

```
{
  ``sentence_num``: 41,
  ``sentence``: ``Samples were fixed in 10% buffered formalin
and embedded in paraffin.``,
  ``measurements``: [
    {
      ``number``: ``10``,
      ``unit``: ``%``,
      ``related``: [
        {
          ``Samples``: []
        },
        {
          ``formalin``: [``buffered``]
        }
      ]
    }
  ]
}
```

pattern parsing (used to identify related words) originates at measurement units, units from the ground truth data were fed to the system rather than relying on Grobid Quantities to provide these values. We assume Grobid Quantities can be incrementally improved with additional labeled values and model training. Although generating this additional training data was outside the scope of our experiment, our experience with Grobid Quantities suggests that domain-specific labels and training is necessary to achieve viable levels of precision and recall for measurement value and unit extraction.

### 5.3 Scoring

Marve parses each of the 489 sentences individually. If one or more measurements are in a sentence, Marve’s extraction of related words for each measurement is compared to the corresponding measurements related words in the labeled data. Because modifiers and descriptors are relatively straightforward to extract for an already-identified related word, they were not considered in the evaluation (e.g. “buffered” in Table 4). For instances where Marve’s related word extractions match the related entities in the labeled data (e.g. “Samples” and “formalin” in Table 4), a true positive is recorded for each matched entity. A false positive is recorded for each extracted related word without a match in the labeled evaluation data. A false negative occurs when an entity from the labeled data can’t be matched to the related words extracted by Marve. Lastly, the count of true negatives – unrelated words that weren’t extracted – was deemed excessive for the experiment.

### 5.4 Results

Marve’s precision, recall, and F-score were evaluated for the two datasets and are shown in Table 6. Because Marve is rule-based system rather than a generalized statistical model, the recall metric indicates the extent to which measurement language follows concrete rules rather than how well a given model represents the data. As long as the rules generalize and are relatively concise, this type of system is attractive for its speed and transparency. Our

**Table 5: Experiment Results - Confusion Matrices**

	Predicted Negatives	Predicted Positives
Combined		
Negatives	n/a	55
Positives	115	225
Journal		
Negatives	n/a	36
Positives	84	143
News		
Negatives	n/a	19
Positives	31	82

recall results imply that a rule-based system such as Marve can identify words and entities related to measurements with high fidelity. While some recall error is expected because language is varied and often misused, these results understate Marve’s recall. Similar to the findings in ClausIE’s experiments, our preliminary analysis suggests that a significant portion of recall error resulted from incorrect dependency parsing rather than the occurrence of undefined patterns. As these systems continue to improve, Marve’s recall will improve as well.

Incorrect dependency parsing seems to be the primary source of error in the precision of extractions, as we would expect precision to be nearly perfect aside from these errors. There are some dependency patterns with ambiguity, where they could or could not indicate words related to a measurement. For example, for the sentence ending, “...area of approximately 1300 kmexp 2 (Sanchez, Martfinez-Fernandez, Scaini, & Perez-Gutierrez, 2012),” normally the content in parenthesis is directly related to the measurement. But in this case, this content is a citation that could apply to the entire sentence or various portions of it. These false positives are not as frequent as occurrences of grammatical variety or grammatical errors that cause a related word to be missed. This leads us to believe that precision error is a better indication of dependency parsing error.

It isn’t surprising that precision and recall were better for the New York Times articles. Compared to scientific publications, sentence patterns in news articles are simpler. The use of special characters, references, diverse punctuation, and domain specific lexicon that can fool a dependency parser is far less common in news articles. Also, Stanford CoreNLP’s English parser is trained on the Penn Treebank which contains a large share of Dow Jones Newswire stories and a much smaller portion of scientific abstracts from the Department of Energy (DOE) [15]. Although there were performance differences across individual data sources, more data is needed to make generalizations about individual domains (e.g. performance in remote sensing versus medicine). The nature of the Penn Treebank data doesn’t provide any clues either, as the abstracts from the DOE were from a variety of scientific domains.

These results indicate that Marve is a sound approach to extracting words forming the context around a measurement. Performance will improve as Marve’s language pattern rules are further scrutinized, extended, and refined, and advances in underlying NLP approaches will also lift performance.

**Table 6: Experiment Results - Evaluation Metrics**

Metric	News	Journal	Combined
Precision	81.2%	79.9%	<b>80.4%</b>
Recall	72.6%	63.0%	<b>66.2%</b>
F-Score	76.6%	70.4%	<b>72.6%</b>

## 6 APPLICATIONS

### 6.1 Opportunities for HypsIRI

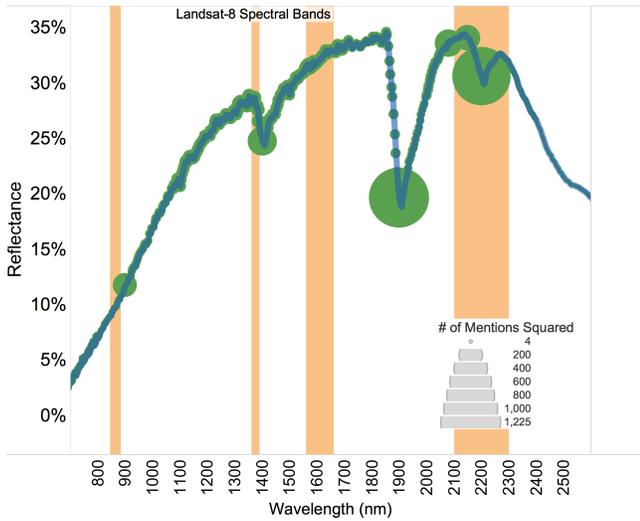
Marve extractions enabled deeper analysis into the potential of the HypsIRI mission mentioned in section 2. Within our corpus of remote sensing-related journal articles, we were able to extract measurement values and units with improved precision and recall. Extractions of related words and entities allowed us to group measurement types with more confidence. They also created opportunities to link semantic publication data to other structured scientific data – an area of research largely unexplored in Earth Science.

Extracted contextual words often contained geophysical features related to certain measurements. Geophysical features can include types of vegetation, soils, minerals, rocks, water, and man-made features, and they are often targeted for measurement by earth science missions. For such missions, the ability to classify certain features is essential to meeting the scientific objectives of the mission. One common way to do this involves analyzing a feature’s reflectance, which varies across electromagnetic wavelengths to form a signature []. These signatures then contain unique combinations of inflection points, helping models to distinguish between different earth-based features. Given their importance to mission science objectives, we were curious to explore the discussion around inflection points for certain geophysical variables.

To understand discussion around inflection points for certain variables we first needed to find measurements extractions referring to the portions of the electromagnetic spectrum. This was straightforward – measurement units of microns or nanometers were strong positive indicators of a spectral reference. One alternative type of measurement, spectral resolution, is also generally referred to using nanometers but in remote sensing these values are usually less than 100 nanometers.

The next step was joining these measurements with associated reflectance signatures. The ASTER spectral library contains over 2,300 spectra of a variety of materials, several of which were present in the related words extracted with measurements [5]. One such example can be seen in figure 2, which shows the reflectance signature of a sample of brown to dark brown clay. There is a large inflection point at 1,900 nm where there also are several mentions of “clay” in association with this wavelength (e.g. “1900 nm” or “1.9  $\mu\text{m}$ ” or “1800-1900 nm”). This figure also shows mentions around wavelengths that aren’t obvious inflection points. They may warrant discussion for other reasons, such as their importance in a specific scientific application.

In addition to scientific objectives, accurate remote sensing of clay types has tangible downstream benefits to various commercial industries. For example, Polyvinyl chloride (PVC), the third most-produced placed plastic polymer, is often combined with clay



**Figure 2: The reflectance of a sample of brown to dark brown clay is indicated by the blue line. The green circles represent the number of extractions that indicated “clay” was related to a specific wavelength or range of wavelengths. The bands represent the spectral coverage of the Landsat-8 satellite. This chart supports the intuitive idea that discussion about reflectance of geophysical features would be centered around inflection points.**

minerals to form nanocomposites used as plasticizers [18] [20]. Specific compositions of clay mineral deposits are widely used in the production of ceramics [16]. And clay composition information helps in understanding absorption properties, which can be used to manage water irrigation more efficiently [7]. As previously discussed, understanding the extent to which remote sensing can support these applications first involves gathering measurement requirements. Once these are defined, assessment of existing satellites and airborne instruments is necessary to identify needs and opportunities (for HypsIRI in this case). Again, this can be accomplished through integration of structured data with Marve extractions. Figure 2 shows bands indicating the spectral coverage of Landsat-8, a satellite developed by NASA and the U.S. Geological Survey (USGS) and launched in 2013 [17] [12]. The pronounced inflection point at 1,900 nm isn’t captured by Landsat-8, which could represent an opportunity for HypsIRI. If other publicly-sponsored satellites also are unable to capture this wavelength and HypsIRI does, HypsIRI would be uniquely positioned to support scientific and industrial applications relying on accurate classification of types of clay.

True understanding of visualizations like figure 2 require deeper research. And precise quantification of the impact of additional data provided by HypsIRI is impossible. However, the automatic collection and integration of semantic data with existing structured scientific data represents a large step forward identifying where further research is warranted and accelerating similar analysis.

## 6.2 Knowledge Base Construction

The construction of a measurement knowledge base for remote sensing publications would be of great value to researchers and policy-makers. For example, in the *Remote Sensing of Environment* paper used in our experiment the authors write, “Although there is more and more information about these soil water parameters, they are not usually included in standard soil databases. For that reason, researchers sometimes have simply used soil parameter data published in the literature.” A structured repository of measurement information would allow researchers to better explore scientific results, experimental designs, instrument specifications, and general discussion around specific measurement types and their relationships to time, locations, organizations, and other domain specific entities. The value and feasibility of constructing similar knowledge bases has been demonstrated by a DeepDive system, PaleoDeepDive that has processed over 300,000 scientific documents in an effort to replicate the manually curated Paleobiology Database (PBDB). This database contains hundreds of thousands of taxonomic fossil names and attributes manually entered by researchers who have sifted through scientific documents over the last two decades. PaleoDeepDive was able to re-create PBDB with greater than double the recall of humans and roughly equal precision.

Instead of taxonomic fossil information, an Earth Science knowledge base would include metrics like classification accuracies for various geophysical variables, instrument specifications like revisit rate, spectral resolution, or swath, and “ground-truth” data used in evaluating remotely-sensed data. Marve significantly reduces the manual effort needed to create such a knowledge base and can be easily integrated into DeepDive, which can add non-measurement-based extractions and also help categorize measurements and their relationships with other entities. DeepDive has previously been used to derive relationships between measurements and related words or entities (e.g. PaleoDeepDive), but users are left to their own devices to extract measurements and possible related words and entities. They also need to write features used by the inference engine to identify and classify relationships between extractions. Marve automates these extractions and can provide DeepDive developers with valuable features (e.g. related context) for further classifying measurements and their surrounding relationships.

Earth Scientists are the most obvious beneficiaries of an Earth Science knowledge base, but several other decision-makers would benefit as well. For example, Resources for the Future (RFF) is leading a NASA-funded consortium, Valuation of Applications Benefits Linked with Earth Science (VALUABLES), which is studying the socioeconomic benefits of Earth observations. They are utilizing the Value of Information (VoI) framework to quantify value as a function of decision-maker uncertainty, what’s at stake, the cost to make use of information, and the cost of substitutes (e.g. space-derived data versus airborne data) [13]. An Earth Science knowledge base could inform all of these areas, helping policy-makers and stakeholders evaluate past and future missions.

## 7 FUTURE WORK

### 7.1 Expanding Experimentation

Research and applications involving measurement extraction are limited. Marve creates second-order extractions (i.e. extractions

built using other extractions) using first-order extractions that are either new (Grobid Quantities) or have progressed significantly in recent years (Word dependencies, POS tagging). As a result, publicly-available labeled datasets don't exist for evaluating Marve. The evaluation dataset used in our experiment is relatively small, and we hope to extend this dataset and employ a linguist to review labels.

Marve also relies on three different types of first-order extractions: CoreNLP's POS tags, CoreNLP's word dependencies, and Grobid Quantities' measurement values and units. We were able to easily exclude Grobid Quantities as a source of error in our experiment by using the labeled evaluation measurement value and units rather than output from Grobid Quantities. However, Marve is more tightly coupled with CoreNLP and CoreNLP errors could not be isolated in the experiment without hand labeling of POS tags and word dependencies. This process is tedious and requires lingual expertise to perform accurately and consistently. While we have focused on measurement-rich scientific publications in our development and applications of Marve, we plan to explore its performance on syntactically labeled data such as the Penn Treebank [21]. Although measurements will be more sparse, experiments with such data sources will allow for Marve to be evaluated independent of first-order extractions.

We are also interested in understanding the performance of Grobid Quantities at different levels of training and customization. Although generating additional training data was outside the scope of our experiment, we are curious how well a Grobid Quantities model trained on a pre-existing set of labeled data would generalize. This will allow practitioners to weigh the costs and benefits of domain-specific labels and training by understanding Grobid Quantities' out-of-box performance on unseen data. Then, as we expand evaluation data for Marve, we hope to also work with groups interested in measurement extraction to generate training examples for Grobid Quantities in various domains. This will help us understand the ceiling for precision and recall for extraction of measurement values and units.

COULD TALK ABOUT CLEAR EARTH AS EXAMPLE OF PARTNERSHIP

## 7.2 Extending Marve

Grobid Quantities and CoreNLP both have difficulty handling long inputs. For Grobid Quantities, this is addressed by using CoreNLP's sentence splitter to pass individual sentences to Grobid Quantities. However, users must perform chunking for longer inputs to CoreNLP and consequently Marve. We plan to automate this data preparation by integrating paragraph-level chunking for various input formats. This will include splitting on <p> tags for XML or HTML input or

n characters in extracted publication content. We plan to eventually extend our pipeline to accept inputs to raw documents using Apache Tika. Integration of Tika will allow us to extract textual content from most document types, which can then be passed to our chunker before being fed to the existing Marve system [].

In addition to extending pre-processing capabilities, several approaches have been identified to improve the usefulness of extractions. For example, Marve doesn't leverage the full extent

of CoreNLP. Incorporating CoreNLP's NER tagging would allow words related to measurements to be categorized into groups like *person*, *location*, or *organization*. With this information in Marve's output, users will have richer information to group and understand measurements (e.g. "Which measurements were related to HyspIRI?"). Another component of CoreNLP, coreference resolution, can also be used to disambiguate pronouns included in extractions.

### N-GRAMS

Increased semantic information embedded in Marve extractions increases the potential for automatic classification of measurements. While Marve represents a large step forward in the collection of semantic measurement information, the burden is on the user to make use of this information, which will involve grouping or classifying measurements in almost all cases. DeepDive addresses this problem by allowing users to write "labeling functions," which provide the system with features used to classify different types of relationships. We plan to explore further integration of Marve into DeepDive and its new successor, Snorkel, while also exploring unsupervised approaches to measurement grouping. We view providing a means for automatically or semi-automatically classifying measurements as an important step in Marve's development.

## 8 DISCUSSION AND CONCLUSIONS

We propose a baseline method for contextual measurement extraction, a sub-area of information extraction that has been largely unaddressed in the research community. Our initial findings are positive, and as we continue to improve Marve and apply it in new ways we hope to spur others to explore new approaches and employ them for their own purposes. Semantic measurement information is inherently richer than raw data – someone found it worthy of discussion. And as the world becomes increasingly inundated with textual data, Marve and other related approaches will help us find relevant information and develop a broader understanding of our domains. We view Marve as an opportunity to expedite scientific research and inform scientific investment, two areas essential to encouraging innovation and demonstrating the importance of science to society.

## ACKNOWLEDGMENTS

This effort was supported in part by JPL, managed by the California Institute of Technology on behalf of NASA, and additionally in part by the DARPA Memex program. The authors would like to thank Jason Hyon, Paul Ramirez, Dr. David Thompson, Dr. Diane Evans, Dr. Randy Friedl, and Dr. Duane Waliser for their support and feedback.

## REFERENCES

- [1] Milan Agatonovic, Niraj Aswani, Kalina Bontcheva, Hamish Cunningham, Thomas Heitz, Yaoyong Li, Ian Roberts, and Valentin Tablan. 2008. Large-scale, parallel automatic patent annotation. In *Proceedings of the 1st ACM workshop on Patent information retrieval*. ACM, 1–8.
- [2] Alan Akbik and Alexander Löser. 2012. Kraken: N-ary facts in open information extraction. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, 52–56.
- [3] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042* (2016).

- [4] Hidir Aras, René Hackl-Sommer, Michael Schwantner, and Mustafa Sofean. 2014. Applications and Challenges of Text Mining with Patents.. In *IPaMin@KONVENS*.
- [5] AM Baldrige, SJ Hook, CI Grove, and G Rivera. 2009. The ASTER spectral library version 2.0. *Remote Sensing of Environment* 113, 4 (2009), 711–715.
- [6] Hannah Bast and Elmar Haussmann. 2013. Open information extraction via contextual sentence decomposition. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*. IEEE, 154–159.
- [7] Mohamed R Berber, Inas H Hafez, Keiji Minagawa, Masami Tanaka, and Takeshi Mori. 2012. An efficient strategy of managing irrigation water based on formulating highly absorbent polymer–inorganic clay composites. *Journal of Hydrology* 470 (2012), 193–200.
- [8] Danqi Chen and Christopher D Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks.. In *EMNLP*. 740–750.
- [9] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. A framework and graphical development environment for robust NLP tools and applications.. In *ACL*. 168–175.
- [10] Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 355–366.
- [11] Christine M Lee, Morgan L Cable, Simon J Hook, Robert O Green, Susan L Ustin, Daniel J Mandl, and Elizabeth M Middleton. 2015. An introduction to the NASA Hyperspectral InfraRed Imager (HyspIRI) mission and preparatory activities. *Remote Sensing of Environment* 167 (2015), 6–19.
- [12] Charlie LLOYD. 2013. Landsat 8 Bands. (2013). <http://landsat.gsfc.nasa.gov/landsat-8/landsat-8-bands/>
- [13] Molley K Macauley and others. 2005. *The Value of Information: A background paper on measuring the contribution of space-derived earth science data to national resource management*. Resources for the Future.
- [14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. <http://www.aclweb.org/anthology/P/P14/P14-5010>
- [15] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* 19, 2 (1993), 313–330.
- [16] Abdou Mbaye, Cheikh Abdoul Khadir Diop, Benaïssa Rhouta, JM Brendle, François Senocq, Francis Maury, and Dinna Pathé Diallo. 2012. Mineralogical and physico-chemical characterizations of clay from Keur Saër (Senegal). *Clay Minerals* 47, 4 (2012), 499–511.
- [17] NASA. 2013. Landsat 8 Overview. (2013). <http://landsat.gsfc.nasa.gov/landsat-8/landsat-8-overview/>
- [18] R Patt, O Kordsachia, R Süttinger, Y Ohtani, JF Hoesch, P Ehrler, R Eichinger, H Holik, U Hamm, ME Rohmann, and others. 2002. Ullmannfis encyclopedia of industrial chemistry. (2002).
- [19] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, and others. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 523–534.
- [20] Tatiane F Silva, Bluma G Soares, Stephane C Ferreira, and Sebastien Livi. 2014. Silylated montmorillonite as nanofillers for plasticized PVC nanocomposites: Effect of the plasticizer. *Applied Clay Science* 99 (2014), 93–99.
- [21] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn treebank: an overview. In *Treebanks*. Springer, 5–22.