Mars Target Encyclopedia:
Rock and Soil Composition Extracted from the Literature

Kiri L. Wagstaff,[1] Raymond Francis,[1] Thamme Gowda,[1,2*] You Lu,[1]
Ellen Riloff,[3] Karanjeet Singh,[1*] Nina L. Lanza[4]
[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, {firstname.lastname}@jpl.nasa.gov
[2]Information Sciences Institute, University of Southern California, Marina Del Rey, CA 90292, tg@isi.edu
[3]School of Computing, University of Utah, Salt Lake City, UT 84112, riloff@cs.utah.edu
[4]Los Alamos National Laboratory, Los Alamos, NM 87545, nlanza@lanl.gov

Abstract

We have constructed an information extraction system called the Mars Target Encyclopedia that takes in planetary science publications and extracts scientific knowledge about target compositions. The extracted knowledge is stored in a searchable database that can greatly accelerate the ability of scientists to compare new discoveries with what is already known. To date, we have applied this system to ~6000 documents and achieved 41–56% precision in the extracted information.

Introduction

Scientists everywhere are overwhelmed by the stream of new information that is published by their disciplines' conferences, workshops, and journals. It is increasingly difficult to come up to speed in a new area and to stay current with the latest discoveries. In planetary exploration, new discoveries can occur each time new data is transmitted. For example, our rovers on Mars have sent back compositional data for thousands of individual targets (e.g., rocks, soils), and some of those observations have transformed our understanding of past environments on the planet (Grotzinger et al. 2014).

To interpret new observations correctly, it is necessary to be able to compare them with what is already known. For example, if we observe high manganese content at a particular location, we want to know whether it is consistent with previous observations or it indicates an anomalous new discovery. However, no central database exists in which planetary scientists can quickly make that determination.

We have created a system called the Mars Target Encyclopedia (MTE) that uses information extraction methods to analyze planetary science publications and identify stated compositional relationships between Mars surface targets and elements or minerals. The extracted information is stored in a searchable database that allows users to ask questions such as "Which targets contain hematite?" or "What is known about target Dillinger?" It also enables entirely new kinds of information visualization, such as a map display of all locations where the Mars rover Curiosity has detected hematite. Ultimately, the MTE may serve as a resource to inform decisions about the next steps in Mars exploration.

In this paper, we describe the MTE system and its component technologies, the empirical performance of the system on labeled data, and results from a large-scale evaluation on ~6000 documents. The MTE is currently being integrated into a public website called the PDS (Planetary Data System) Analyst's Notebook for Mars scientists and the public to access. The automated pipeline can be used to ingest and analyze new publications as they become available.

Related Work

A variety of text analysis methods exist for extracting information from text. Some methods focus on extracting meta-data such as the document title, authors, and publication venue or analyzing and linking citations between papers (Ronzano and Saggion 2016). However, understanding the content of a scientific publication requires a deeper analysis. Information extraction (IE) of this nature is generally broken into two steps: (1) named entity recognition or concept extraction, to identify references to people, locations, concepts, etc., and (2) relation extraction, to identify relationships between pairs of entities (Mooney and Bunescu 2005). Many of the recent advances in IE have been motivated by problems from the biomedical research world, such as the desire to identify protein-protein interactions (Tikk et al. 2010; Bui, Katrenko, and Sloot 2011) or chemical-protein and chemical-disease relations (Krallinger et al. 2017). Tsutsui, Ding, and Meng (2016) used topic modeling and open IE, which does not require the prior identification of entities, to build a knowledge database about Alzheimer's disease.

To date, little such work has been done in the domain of planetary science. The closest existing work is the geology-based GeoDeepDive project, which performs text data mining on scientific publications about (Earth) rock formations and stratigraphy (Zhang et al. 2013). By applying and extending information extraction methods to planetary science publications, we have the opportunity to benefit an entirely new population of scientists, researchers, and interested members of the public.

Machine Learning for Information Extraction

The Mars Target Encyclopedia (MTE) is an information extraction system that takes in scientific publications in PDF format and extracts knowledge that is useful to scientists

[*]This work was done when the authors were at the Jet Propulsion Laboratory.

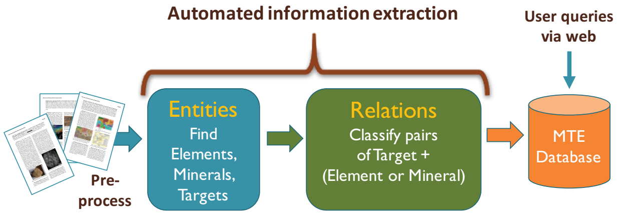

Automated information extraction
User queries via web
Pre-process
Entities
Find Elements, Minerals, Targets
Relations
Classify pairs of Target + (Element or Mineral)
MTE Database

Figure 1: Mars Target Encyclopedia processing pipeline.

studying the planet Mars. We focus on the extraction of information about targets (e.g., rocks, soils) identified by the ChemCam instrument on the Curiosity rover. ChemCam uses a laser spectrometer to obtain compositional spectra from up to seven meters away from a given target. The resulting spectra can be analyzed to identify individual elements within the target (Maurice and 70 others 2012). As of sol 1159 of the Curiosity rover's mission, ChemCam had observed more than 1100 distinct targets. New discoveries about these targets are published in a variety of planetary science conference and journal venues.

The MTE is composed of four modules: preprocessing, named entity recognition (NER), relation extraction (RE), and database updates (see Figure 1).

Document Preprocessing

To prepare the documents for information extraction, the MTE first extracts the text content from each PDF document. We use the Apache Tika parser (Mattmann and Zitting 2011) to convert the source PDF files into UTF-8 format text to preserve accented characters and mathematical symbols. Next, the MTE creates a copy of the text content in which the References section is omitted. We defined a regular expression to identify the References section. This step helps the NER module avoid spurious detections in the titles or author names of cited publications.

Named Entity Recognition

The named entities of greatest relevance for the MTE are elements (e.g., "iron," "Mg"), minerals (e.g., "plagioclase," "hematite"), and ChemCam targets. The periodic table provides a comprehensive list of elements, and we employed a list of 5228 minerals provided by the International Mineralogical Association (the May 2017 release[1]).

Identifying Mars surface targets is more challenging, as they follow no standard naming convention. Further, target names are fundamentally ambiguous as they are borrowed from Earth locations or people. A sampling of the names hints at the challenge of accurately detecting them: "Dunkirk", "Ithaca", "Jake", "Old woman", "Pistol". We have a starting list of target names[2] that was published by the ChemCam science team, but it is not fully curated, and we have found it to be incomplete with respect to the literature. In addition, we found that authors continually invent new spelling variants and abbreviations for target names that require more than a simple list lookup.

To address the challenge of recognizing all three entity classes reliably, we employed a machine learning approach. We trained a custom Named Entity Recognizer using the Stanford CoreNLP NER system (Finkel, Grenager, and Manning 2005). This system trains a Conditional Random Field sequence model to assign class labels to entities within new documents. We provided manually labeled documents with examples of the Element, Mineral, and Target classes to train our custom model. We also employed the "gazette" capability to provide lists of known terms. This is particularly valuable for large semantic classes (like Mineral or Target) in which terms exist that might never appear in the training corpus. The gazettes that we used include the periodic table (Elements), the IMA list (Minerals), and the ChemCam observation table (Targets) as mentioned above.

Relation Extraction

Once the entities are identified within the text, the MTE analyzes them to determine which ones have a compositional relationship (i.e., textual evidence that a given Target contains a given Element or Mineral). We trained a relation classifier using the jSRE (Giuliano, Lavelli, and Romano 2006) relation extraction tool. It uses an SVM classifier to predict whether a relationship exists for two entities, using only shallow parsing. jSRE provides SVM kernels that operate on local context, global (sentence-wide) context, or a combination of both. In applying this method to biomedical publications, the authors found that most of the performance came from the global kernel features.

MTE Database

For each document, the MTE database stores document meta-data (e.g., title, author, publication venue), the preprocessed text content, and the extracted entities and relations. To provide full traceability, it also stores the parsed sentence from which the relation was extracted and a link to the original PDF. Users can immediately see the source context and decide whether they would like to access the full document for more information.

To support fast retrieval of documents that contain entities or relations specified in a given search query, we constructed inverted indices for the entities and relations using Apache Solr[3]. Apache Solr is an open source search platform powered by Apache Lucene that provides near real-time search and document retrieval.

Experimental Results

We developed and evaluated the MTE using a collection of scientific papers that were published over three years of the Lunar and Planetary Science Conference (LPSC). These papers are publicly accessible from the LPSC websites (e.g., https://www.hou.usra.edu/meetings/lpsc2014/).

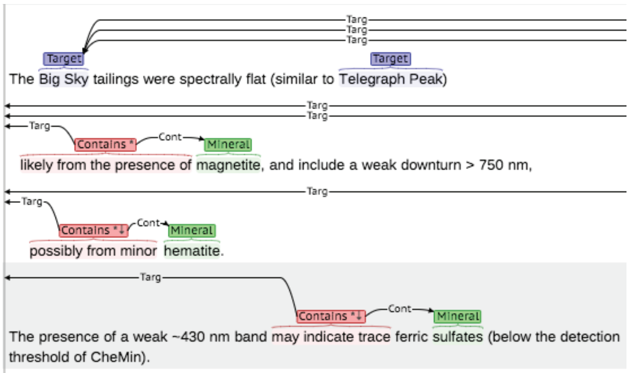
[1] http://nrmima.nrm.se/imalist.htm

[2] http://pds-geosciences.wustl.edu/msl/msl-m-chemcam-libs-4_5-rdr-v1/mslccm_1xxx/document/msl_ccam_obs.csv

[3] http://lucene.apache.org/solr/

Table 1: Manual annotations for LPSC documents.

Annotation	2015 (62 docs)	2016 (55 docs)	Total (117 docs)
Element	1195	1029	2224
Mineral	748	708	1456
Target	566	347	913
Contains	434	262	696
Total	2943	2346	5289



The Big Sky tailings were spectrally flat (similar to Telegraph Peak)

likely from the presence of magnetite, and include a weak downturn > 750 nm,

possibly from minor hematite.

The presence of a weak ~430 nm band may indicate trace ferric sulfates (below the detection threshold of CheMin).

Figure 2: Excerpt from document lpsc16-1155 showing compositional annotations created with the brat web annotation tool.

Corpus

Our corpus consists of two-page extended LPSC abstracts in PDF format. We selected 62 documents from LPSC 2015 and 55 documents from LPSC 2016 that mentioned "ChemCam", and used the brat annotation tool (Stenetorp et al. 2012) to manually label entities within these documents (see Table 1). This data set contains thousands of annotations, which are available here: https://doi.org/10.5281/zenodo.1048419. We estimate that it took an average of 30 minutes to annotate each document, or a total of more than 58 hours of labor for the full corpus.

The annotated relationships ranged from simple (e.g., a pattern such as "X contains Y" within a sentence) to complex (e.g., relationships that crossed sentence boundaries or involved pronouns like "it" and other anaphora). Figure 2 shows an excerpt from one document that contains several statements about the composition of the target Big Sky. The vocabulary used to indicate a compositional relationship varies, and the final relationship crosses a sentence boundaries.

Named Entity Recognition Results

The Named Entity Recognizer operates on individual words or tokens. We used the 2015 documents for training and divided the 2016 documents into validation ($n = 20$) and testing ($n = 35$) sets. As shown in Table 2, the baseline approach of employing the known lists of elements, minerals, and targets achieved an F1 score of 0.76. Training a basic NER classifier using the CoreNLP system yielded an improved F1 score of 0.805. Virtually all of the improvement came from increased precision (from 0.83 to 0.95). Recall was highest (0.84) for the Element class, as expected; it was 0.73 for Minerals and only 0.28 for Targets. The Target class is the most difficult one to recognize due to the lack of a naming convention and ambiguous names. In addition, the Target class grows much faster than the set of known elements or minerals, so there will always be new targets in future documents that never appeared in the training set.

Table 2: Named entity recognition performance on LPSC 2016 test documents. The best result for each metric is shown in bold.

	Prec.	Recall	F1
Baseline: Lists only	0.831	0.699	0.760
CoreNLP NER trained on:			
LPSC 2015	**0.948**	0.700	0.805
LPSC 2015 + gazettes	0.945	**0.777**	**0.853**

However, we were able to improve NER recall as well by including the gazettes as described above. These term lists augment the manually labeled documents and provide relevant domain knowledge. With the gazettes, the F1 score increased to 0.853 by boosting recall to 0.777. Recall for the Target class, in particular, more than doubled, to 0.67.

Relation Extraction Results

The decision about whether or not a relation exists is made for a given pair of entities. Processing all possible pairs of entities in a document would be infeasible (and likely unnecessary). For simplicity, we adopted the strategy used in previous work (Giuliano, Lavelli, and Romano 2006) of generating all pairs of entities that occur within a single sentence. We used CoreNLP's sentence splitter to divide the corpus into sentences and the NER model trained above to identify entities. For each (Target, Element) or (Target, Mineral) pair, we generated a jSRE example that encoded the sentence content. If the pair of entities was connected by a relation in the manual annotations, we gave the example a positive label; otherwise, we gave it a negative label.

To simulate how the system would be used in practice, we trained and validated the relation classifier using text from LPSC 2015 and tested it on LPSC 2016. We used the first 42 LPSC 2015 documents for training and the remaining 20 for validation. The number and distribution of the resulting jSRE examples (relationships) are given in Table 3.

Table 3: Number and distribution of relationships between Targets and Elements or Minerals. The number in parentheses is the percentage of positive relationships.

	Element	Mineral	Merged
Train	279 (38%)	150 (41%)	429 (39%)
Validation	93 (27%)	70 (69%)	163 (45%)
Test	111 (37%)	62 (50%)	173 (42%)

Table 4: Relation extraction performance on LPSC 2016 (test) documents. The best result for each metric is shown in bold.

	Precision	Recall	F1
Elements ($n = 111$)			
Baseline: All-yes	0.369	**1.000**	**0.539**
jSRE-Elements	**0.531**	0.415	0.466
Minerals ($n = 62$)			
Baseline: All-yes	0.500	**1.000**	**0.667**
jSRE-Minerals	**0.679**	0.613	0.644
Merged ($n = 173$)			
Baseline: All-yes	0.416	**1.000**	**0.588**
jSRE-Indiv.	0.598	0.447	0.511
jSRE-Merged	**0.640**	0.444	0.525

We trained three different relation classifiers: one on Target-Element relations only; one on Target-Mineral relations only; and one on the merged set. We were curious as to how a specialized model that was trained on less data would compare to a more generic model trained on more data. For each model, we performed a grid search over the jSRE parameters by trying each of the SVM kernels (LC, GC, SL) and window sizes within the set { 1, 2, 5, 10, 15, 20 }. We selected the model parameters that led to the highest performance on the validation set in terms of precision. We found that the max-precision model did not employ the same parameters across the three models. jSRE-Elements and jSRE-Merged used an LC kernel with a window of 5, while jSRE-Minerals used an SL kernel with a window of 5. Notably, the GC kernel that the original authors found to be most powerful for the biomedical domain did not perform well in this corpus.

We found that the individual models ("jSRE-Elements" and "jSRE-Minerals") achieved much higher precision than a baseline approach that always predicted that a relationship was present ("All-yes") (see Table 4). While this baseline always achieves a recall of 1.00 and therefore appears superior in terms of F-measure, this application domain values precision much more than recall. Content included in the MTE must be of the highest reliability, even if this means it is not comprehensive. We also found that the merged model ("jSRE-Merged") out-performed the baseline and the individual models when they were applied to the full (Merged) data set ("jSRE-Indiv.").

Large-scale Evaluation

We collected all LPSC documents that were published in 2014, 2015, and 2016, omitting the training documents from LPSC 2015, and ingested them into the MTE ($n = 5897$).

It would be infeasible to ask humans to manually label all 5897 documents to evaluate our results, so instead we performed a manual review of only the extracted relations. This allows us to measure precision, but not recall. However, as noted above, precision is far more important than recall in this domain, as it captures the true utility of the extracted information when used in practice.

Table 5: Manual review of 817 relations extracted from 5897 documents.

	LPSC14	LPSC15	LPSC16	Total
Correct	55%	57%	29%	41%
Partial	9%	15%	14%	13%
Irrelevant	19%	6%	9%	11%
Wrong	11%	21%	2%	8%
Unsure	6%	0%	47%	28%

The manual review results are shown in Table 5. Our manual reviewer examined each extracted relation and its source sentence to judge the relation as Correct, Partial (e.g., only one word of a multi-word Target name was extracted), Irrelevant (an appropriate extraction from the sentence, but the content was not about Mars), Wrong, or Unsure (the reviewer could not determine whether the relation was correct).

Overall, the fraction of Correct relations was 41%. Performance on the 2016 documents was significantly lower than for the preceding years. Since the system was trained on documents from 2015, it is likely that targets mentioned in 2015 would encompass those discovered in 2014 and 2015, while the documents from 2016 contain many newly discovered targets and therefore present a more difficult generalization task. Many of the Partial relations occurred due to limited support for extracting multi-word entities. This is an area for future improvement.

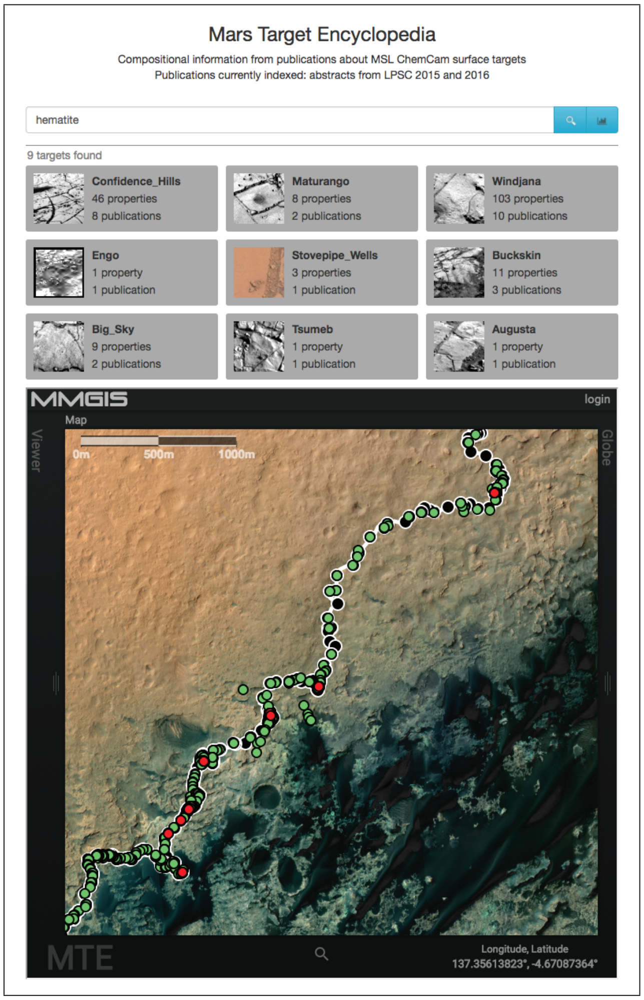
The Irrelevant relations are in some ways quite interesting; there are several relations that express the composition of meteorites that happen to have the same names as (real) Mars targets. The system correctly interpreted the source sentences, but the information does not (strictly) belong in the MTE. For example, the system inferred that "Gibeon" contains "chromite" from this sentence: "Gibeon was found in several studies to have both chromite and daubrelite inclusions." Gibeon is the name of a Mars target and of a meteorite. Disambiguating the two requires more context than a single sentence. Most of the Unsure relations came from tables whose formatting was lost in the conversion from PDF to text. A useful future direction might be to omit table content or to capture its structure in some way, e.g., by using the Tabula tool[4]. If we omit these unparseable sections, we obtain 56% Correct relations, 18% Partial, 15% Irrelevant, and 11% Wrong.

Deployment of the MTE

We created a simple web interface to allow users to query the MTE for information about targets, elements, or minerals. This interface is currently only available inside JPL, but we are in the process of integrating it with a public PDS website as discussed below.

The MTE enables scientists to ask new questions that previously could not be answered. For example, Figure 3 shows the results of a query on "hematite." Nine targets that con-

[4] https://github.com/tabulapdf/tabula



Mars Target Encyclopedia

Compositional information from publications about MSL ChemCam surface targets
Publications currently indexed: abstracts from LPSC 2015 and 2016

hematite

9 targets found

Confidence_Hills
46 properties
8 publications

Maturango
8 properties
2 publications

Windjana
103 properties
10 publications

Engo
1 property
1 publication

Stovepipe_Wells
3 properties
1 publication

Buckskin
11 properties
3 publications

Big_Sky
9 properties
2 publications

Tsumeb
1 property
1 publication

Augusta
1 property
1 publication

Longitude, Latitude
137.35613823°, -4.67087364°

Figure 3: MTE search results for "hematite." Nine results (individual Mars targets) are returned, and a map displays the location of each hit, in red.

tain hematite were returned. The user can click on any target to see the extracted information and sentence excerpts that support the conclusion about the presence of hematite. Below is a map of the Curiosity rover's traverse on Mars, with the locations of the matching targets marked in red. One can immediately see whether hematite is localized or has been identified throughout the mission.

Limitations

The MTE is not comprehensive. There may be compositional information that was never written up in a scientific publication and therefore would not be included in the MTE. Instead, the MTE extracts and indexes only the information that was judged by scientists to be worthy of publication to the scientific community. The MTE leverages and mirrors this selection bias, and its holdings (like the source publications) contain only the most valuable and salient information. This incompleteness is important to convey to the user so that they interpret results correctly.

On the technical side, there are two important limitations to the MTE content. First, the current MTE cannot generate overlapping annotations. For example, the phrase "calcium sulfate" was manually labeled as "calcium" (Element), "sulfate" (Mineral), and "calcium sulfate" (Mineral). However, the MTE's NER model only classifies individual tokens, so it misses the "calcium sulfate" phrase.

Second, the relation extraction module only generates candidate relation pairs within a sentence. In this corpus, 32% of the manually annotated relations cross sentence boundaries. Therefore, the current system cannot yet retrieve those relations. One way to access sentence-crossing relations would be to expand the number of candidate entity pairs to include all pairs within a paragraph. We plan to evaluate that strategy in future work.

Conclusions and Next Steps

This work lies at the intersection of information extraction, machine learning, and planetary science. The MTE uses current IE technology to provide the first database of Mars target compositional knowledge as expressed in the scientific literature. The pipeline is fully automated, and we can employ web crawlers to seek out new (publicly accessible) papers as they are published. We also plan to enable users to submit their own publications for analysis and augmentation of the database.

We are in the process of integrating the MTE's content into the MSL Analyst's Notebook, an interactive web resource for mission scientists and the interested public (Stein and Arvidson 2013). The Analyst's Notebook allows users to browse mission plans, targets discovered, data collected, and summaries of each mission day on Mars. The MTE content will enable the Analyst's Notebook to also connect targets to publications.

The MTE currently contains information about ChemCam targets that was extracted from three years of papers published at the Lunar and Planetary Science Conference. A next logical step is to expand the MTE to encompass targets identified by other instruments on the Curiosity (MSL) rover and other missions such as the Mars Exploration Rovers (Spirit and Opportunity). In addition, we plan to extend the MTE to be able to ingest journal papers that have been published by MSL science team members and the broader community. This information will carry more weight because it comes from peer-reviewed sources; users will be able to restrict their searches to journal papers only, or to obtain all possible results.

The automatic extraction of knowledge from scientific publications can benefit many other areas of scientific inquiry. In addition to biology and medicine, there are opportunities at the intersection between fields such as planetary science and astronomy. For example, there are currently 3,550 confirmed exoplanets (planets outside our solar system) as of November 2, 2017 (NASA 2017). Hundreds of new planet candidates are announced each year in new publications. Desirable properties to extract and store for each planet include its radius, temperature, period, distance from its host star, and more. Compositional relationships exist for elements present in the host star and for constituents in exoplanet atmospheres, with implications for the possible presence of life on other planets. In general, extracting information and relationships into a central, searchable database can help inform new hypotheses and direct future science investigations.

Acknowledgments

This research was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. We thank the Multimission Ground System and Services (MGSS) program and the Mars Science Laboratory project for funding and enthusiastically supporting this work. We also thank Chris Mattmann for his support and partial funding as provided by the DARPA XDATA/Memex/D3M programs and NSF award numbers ICER-1639753, PLR-1348450, and PLR-144562.

References

Bui, Q.-C.; Katrenko, S.; and Sloot, P. M. A. 2011. A hybrid approach to extract protein-protein interactions. *Bioinformatics* 27(2):259–265.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370.

Giuliano, C.; Lavelli, A.; and Romano, L. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, 401–408.

Grotzinger, J. P.; Sumner, D. Y.; Kah, L. C.; Stack, K.; Gupta, S.; Edgar, L.; Rubin, D.; Lewis, K.; Schieber, J.; Mangold, N.; Milliken, R.; Conrad, P. G.; DesMarais, D.; Farmer, J.; Siebach, K.; Calef, F.; Hurowitz, J.; McLennan, S. M.; Ming, D.; Vaniman, D.; Crisp, J.; Vasavada, A.; Edgett, K. S.; Malin, M.; Blake, D.; Gellert, R.; Mahaffy, P.; Wiens, R. C.; Maurice, S.; Grant, J. A.; Wilson, S.; Anderson, R. C.; Beegle, L.; Arvidson, R.; Hallet, B.; Sletten, R. S.; Rice, M.; Bell, J.; Griffes, J.; Ehlmann, B.; Anderson, R. B.; Bristow, T. F.; Dietrich, W. E.; Dromart, G.; Eigenbrode, J.; Fraeman, A.; Hardgrove, C.; Herkenhoff, K.; Jandura, L.; Kocurek, G.; Lee, S.; Leshin, L. A.; Leveille, R.; Limonadi, D.; Maki, J.; McCloskey, S.; Meyer, M.; Minitti, M.; Newsom, H.; Oehler, D.; Okon, A.; Palucis, M.; Parker, T.; Rowland, S.; Schmidt, M.; Squyres, S.; Steele, A.; Stolper, E.; Summons, R.; Treiman, A.; Williams, R.; Yingst, A.; and Team, M. S. 2014. A habitable fluviolacustrine environment at Yellowknife Bay, Gale Crater, Mars. *Science* 343(6169).

Krallinger, M.; Rabal, O.; Loureno, A.; Oyarzabal, J.; and Valencia, A. 2017. Information retrieval and text mining technologies for chemistry. *Chemical Reviews* 117(12):7673–7761.

Mattmann, C., and Zitting, J. 2011. *Tika in Action*. New York: Manning Publications.

Maurice, S., and 70 others. 2012. The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Science objectives and mast unit description. *Space Science Reviews* 170(1):95–166. doi:10.1007/s11214-012-9912-2.

Mooney, R. J., and Bunescu, R. 2005. Mining knowledge from text using information extraction. *ACM SIGKDD Explorations Newsletter* 7(1):3–10.

NASA. 2017. Exoplanet archive. https://exoplanetarchive.ipac.caltech.edu/.

Ronzano, F., and Saggion, H. 2016. Knowledge extraction and modeling from scientific publications. In *Proceedings of the Enhancing Scholarly Data Workshop*, 11–25. Cham: Springer International Publishing.

Stein, T. C., and Arvidson, R. E. 2013. PDS Analyst's Notebook for MSL. In *Proceedings of the 44th Lunar and Planetary Science Conference*, Abstract 1570.

Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; and Tsujii, J. 2012. brat: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Tikk, D.; Thomas, P.; Palaga, P.; Hakenberg, J.; and Leser, U. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Computational Biology* 6(7).

Tsutsui, S.; Ding, Y.; and Meng, G. 2016. Machine reading approach to understand Alzheimers disease literature. In *Proceedings of the Tenth International Workshop on Data and Text Mining in Biomedical Informatics (DTMBIO)*.

Zhang, C.; Govindaraju, V.; Borchardt, J.; Foltz, T.; and and Shanan Peters, C. R. 2013. GeoDeepDive: Statistical inference using familiar data-processing languages. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 993–996.