Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6077–6082
Marseille, 20-25 June 2022

Extracting Space Situational Awareness Events from News Text

Zhengnan Xie[†], Alice Saebom Kwak[†], Enfa George[†], Laura W. Dozal[†], Hoang Van[†], Moriba Jah[‡], Roberto Furfaro[†], Peter Jansen[†]
[†]University of Arizona, USA
[‡]University of Texas at Austin, USA
pajansen@arizona.edu

Abstract

Space situational awareness typically makes use of physical measurements from radar, telescopes, and other assets to monitor satellites and other spacecraft for operational, navigational, and defense purposes. In this work we explore using textual input for the space situational awareness task. We construct a corpus of 48.5k news articles spanning all known active satellites between 2009 and 2020. Using a dependency-rule-based extraction system designed to target three high-impact events – spacecraft launches, failures, and decommissionings, we identify 1,787 space-event sentences that are then annotated by humans with 15.9k labels for event slots. We empirically demonstrate a state-of-the-art neural extraction system achieves an overall F1 between 53 and 91 per slot for event extraction in this low-resource, high-impact domain.
Keywords: information extraction, corpus, space

1. Introduction

Space Situational Awareness (SSA) is the decision-making knowledge required to predict, avoid, operate through, or recover from the loss, disruption, or degradation of space services, capabilities, or activities (Cai et al., 2020). Most governments approach SSA as a physical measurement problem, using global arrays of optical telescopes, radar facilities, radio listening posts, and other assets to directly measure spacecraft properties, orbits, or communications. In addition to being costly, this approach has several known limitations, chief among them that physical measurements (such as that a *spacecraft is spinning*) alone cannot fully capture causal information (e.g. *a spacecraft suffered a gyro failure*) that can be critical when interpreting and responding to events within the space object population (Walls et al., 2016).

In this work we explore extracting space-domain events from non-physical measurements, namely textual input from press releases and news articles. SSA is frequently accomplished by collecting and fusing multi-modal data (Esteva et al., 2020; Le May et al., 2020), where text is a comparatively inexpensive knowledge resource. We hypothesize that text can supply information complementary to physics-based measurements, such as the published causes of spacecraft failures, whether spacecraft operators have planned astronavigation maneuvers, or whether a spacecraft has been safely decommissioned according to United Nations guidelines.

Information extraction (IE) is often reported in high-resource domains where favorable performance is possible using pre-trained models. In contrast, there exist high-importance domains (e.g. nuclear, defense, etc.) that remain under-explored due to low resource availability, and low task transfer from near-domains. Space-domain information extraction poses these challenges in abundance – comparatively few spacecraft

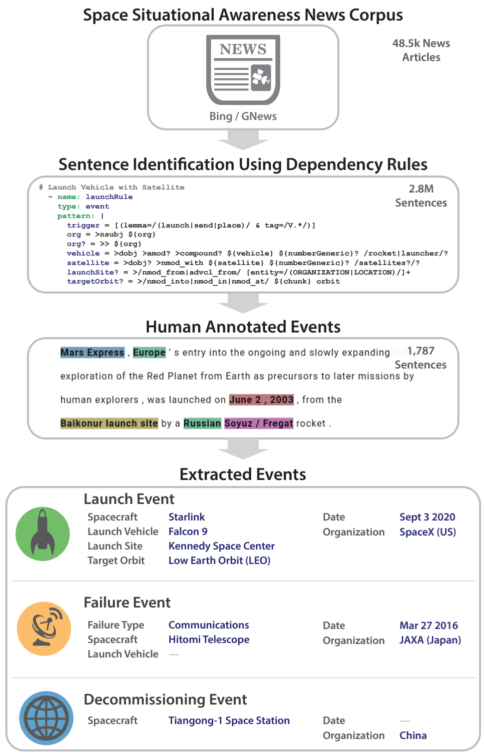

Figure 1: An overview of the proposed system. A corpus of 48.5k space-domain news articles collected from the web serves as input to a sentence-level dependency-based information extraction system, ODINSON, that isolates a shortlist of sentences for human annotation. These are then used to train and evaluate a language-model based event extraction system, which extracts spacecraft *launch, failure,* and *decommissioning* events from text.

have ever existed, and they are meticulously engineered, making launches, failures, and other SSA-relevant events infrequent but highly impactful. By automatically monitoring and extracting this information

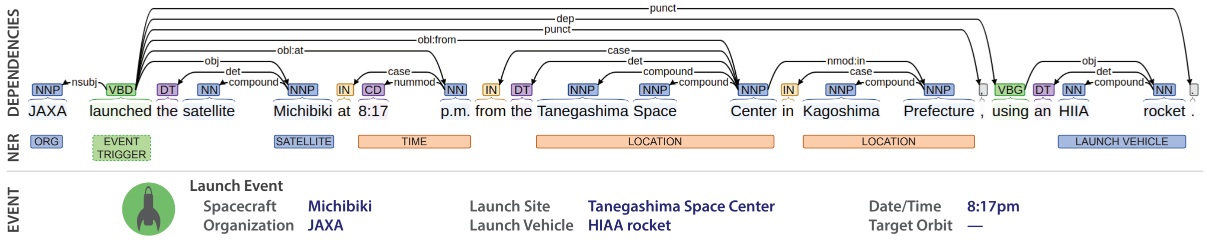

Figure 2: An example of constructing a dependency rule to find candidate *launch* events for manual annotation. Starting with a trigger phrase (here, *launched*), our system traverses specific syntactic dependencies most likely to contain information to fill specific event slots. Both a space-domain *(blue)* and generic CoreNLP *(orange)* NER system assist in identifying slot fillers in high-confidence rules, while low-confidence rules fill slots by extracting chunks from likely dependencies.

from sources available on the web, SSA users can augment existing physical measurements with behavioral information, or provide targets of interest for physical monitoring based on information gathered from text.

In this work we construct the first space-domain event corpus and extraction system, targeted at three demonstrative high-impact SSA events: spacecraft launches, failures, and decommissionings. The contributions of this work are:

1. We collect a large-scale multi-year corpus of 48.5k news articles (containing 2.8M sentences) covering all 4,157 known satellites existing between 2009 through 2020. We pair this with a dependency-rule based extractor that identifies 1,787 sentences with valid launch, failure, or decommissioning events, that we manually annotate with 15.9k labeled tokens representing event slots in each sentence.

2. We empirically demonstrate that in this low-resource high-impact domain, a state-of-the-art language-model-based extraction tool achieves F1 scores between 53 through 91 for extracting different event slots. We highlight common causes of failure, including the large number of distractors in this corpus, in an error analysis.

3. We release our space-domain event corpus and extraction system as open source[1].

2. Approach

An overview of our approach is shown in Figure 1. We approach this SSA task as a multi-slot information extraction problem. Though event information may exist across sentences or entire documents, in this work we focus on events contained within single sentences. First, we assemble a large corpus of space-domain news articles. We then use a dependency-based extraction framework to identify a shortlist of sentences that contain valid events, then manually annotate their event slots as spans in each sentence. Finally, we evaluate the performance of a language-model-based extraction system at identifying these event spans.

2.1. Multi-Slot Event Schema

We focus on extracting three high-impact and comparatively frequent events in the corpus – *spacecraft launches, spacecraft failures, and spacecraft decommissionings*. While IE is often framed as a triple extraction task (Etzioni et al., 2008), here each event schema has between 3 and 6 optional slots to populate with event-relevant information. For example, the LAUNCH event is used to describe the launch of a specific satellite, and may include additional information such as the launch vehicle used (e.g. an *Atlas V*), the launch site (e.g. *Cape Canaveral*), the target orbit (e.g. *geosynchronous*), an organization the launch is on behalf of, and the date the launch occurred. A complete list of the event slots in each schema is included in Table 2.

3. Corpus and Annotation

3.1. Space Domain News Article Corpus

We assembled a list of all known unclassified satellites active at any time between 2009 through 2020 using archived versions of the publicly available UCS satellite database[2], finding a total of 4,157 satellites over this 11 year period. We collected a corpus of space-domain news articles by querying the Microsoft Bing News[3] and GNews[4] APIs for each satellite name (and any alternate names listed by USC), and retrieving the top 100 news articles for each name. As news APIs regularly restrict results to a short temporal period (typically 30 days), and satellite events happen infrequently, collection took place over several years. In total, collection returned 49,908 articles, which were post-processed with Boilerpipe (Kohlschütter et al., 2010) to extract article content while removing advertisements and other extraneous information. News or-

[1] Available at: `https://github.com/cognitiveailab/ssa-corpus`

[2] `https://www.ucsusa.org/resources/satellite-database`

[3] `http://azure.microsoft.com`

[4] `http://gnews.io`

SSA Event	Training			Development			Test		
	Sents.	Tokens *(w/tag)*	Tokens *(total)*	Sents.	Tokens *(w/tag)*	Tokens *(total)*	Sents.	Tokens *(w/tag)*	Tokens *(total)*
Launch	537	5,646	25,855	350	2,890	13,357	350	3,059	12,747
Failure	310	2,748	12,905	63	580	2,656	63	449	2,043
Decommissioning	81	396	3,064	17	75	487	17	68	504

Table 1: Overall statistics for the space situational awareness event corpus across the three events. Sentences represents the number of sentences within a given set. Tagged tokens represents the number of *BIO-tagged* tokens that have a either a *Beginning (B)* or *Inside (I)* tag.

ganizations regularly publish syndicated duplicate articles, or republish updated articles months or years later. To prevent leakage between training and unseen sets, any articles whose unigram cosine similarities exceed a threshold of 0.90 were pooled into the same set. The corpus was divided into training (28.7k document), and unseen(19.8k document) sets, with the most-recently collected articles and newest spacecraft present in the unseen set. The unseen set was evenly split into development and test sets after additional manual filtering described below.

3.2. Identifying Sentences for Human Labelling with Dependency Rules

The space-domain news article corpus contains 2.8M sentences across 48.5k documents. To identify a shortlist of candidate sentences for human labeling that describe space events of interest (launches, failures, and decommissionings), we constructed a set of high-recall dependency-based extraction rules using the ODINSON Information Extraction language (Valenzuela-Escárcega et al., 2020). ODINSON allows expressing extraction rules as a combination of syntactic dependencies, named entities, part-of-speech tags, and surface forms, and processes documents approximately 5 orders of magnitude faster than comparable approaches (Wang et al., 2018; Valenzuela-Escárcega et al., 2016), scanning and extracting sentences in our nearly 50k document corpus in approximately 4 minutes on inexpensive desktop hardware. Articles were indexed using the ODINSON indexer, where Stanford CoreNLP (Manning et al., 2014) supplied tagging, lemmatization, and dependency parsing.

3.2.1. Named Entity Recognition

To support rule-based extraction with ODINSON, we constructed a named entity recognition system for domain-specific slot fillers, including SPACECRAFT (such as *Hubble Space Telescope*), LAUNCHVEHICLES (such as *Falcon 9*), and LAUNCHSITES (such as *Baikonur Cosmodrome*). For open-domain entities, we use CoreNLP entities directly (e.g. DATE), or a combination of both CoreNLP and our space-domain NER system (e.g. ORGANIZATION).

3.2.2. Rule Authoring

Rules were constructed through an iterative process of querying the training set for high-frequency n-grams that might serve as ODINSON trigger phrases associated with each of the 3 events (e.g. *launched, failure,* or *decommissioned*), then authoring slot-extraction rule components based on traversing key syntactic dependencies attached to the trigger phrase. An example dependency tree and associated event is shown in Figure 2. To balance precision and robustness, we constructed 35 high-confidence rules that required NER matches for all slots, while 32 lower-confidence backoff rules allow matching unknown entities or generic NP-chunks for one or more slots, increasing recall for (for example) newer or unknown spacecraft at the expense of precision. In total, 67 rules were authored across all 3 events.

3.2.3. Manual Filtering

The rule-based system identified 4346 *launch*, 824 *failure*, and 218 *decommissioning* candidate sentences from the corpus of 2.8M sentences. Manual inspection showed that 91% of launch sentences and 53% of both failure and decommisioning sentences contained a valid event, while others contained unrelated information (e.g. the launch of a new product) and were filtered out. Due to the large number of events, the *launch* set was randomly subsampled to 30% of its original size for human labeling. A total of 1,787 sentences progressed through this stage, for subsequent labeling by human annotators.

3.3. Annotating Space Events

Annotation took the form of a span-labelling task, where for a given sentence an annotator would highlight relevant spans of text (e.g. *"Hubble Space Telescope"*), and select an appropriate event slot label (e.g. SPACECRAFT). A single sentence can mention more than one event (for example, two satellites being launched), and in these cases all relevant slot fillers are annotated for both events.[5] We used LightTag[6]

[5]Sentences without mention of a specific SPACECRAFT (for launches and decommissionings) or a specific SPACECRAFT or LAUNCHVEHICLE (for failures) were removed from consideration during the manual filtering step.

[6]`www.lighttag.io`

Event and Slot	Development Pr	Re	F1	N	Test Pr	Re	F1	N	Example
Launch Event				*350*				*349*	
SatelliteName	83	89	86	540	85	89	87	528	Hubble Space Telescope
LaunchVehicle	89	89	89	141	86	92	89	170	Space Shuttle STS-32
LaunchSite	86	97	91	88	89	88	88	137	Cape Canaveral
TargetOrbit	62	84	71	25	89	82	74	38	Low-Earth Orbit
Failure Event				*63*				*63*	
SatelliteName	76	87	81	70	69	87	77	53	Telkom-3
LaunchVehicle	80	84	82	38	88	94	91	31	Proton-M
FailureType	41	55	47	65	48	66	56	50	Launch or power failure
Decommissioning Event				*17*				*17*	
SatelliteName	61	85	71	20	44	67	53	18	NOAA-19
Generic Slots Across Events									
Organization	66	81	73	349	71	82	76	375	NASA
Date	83	91	87	434	84	88	86	390	April 4, 1990

Table 2: Overall performance for the language model-based extraction system over the three event schema (*spacecraft launches, failures, and decommissionings*), broken down by event slot. Generic slots can optionally occur across all three event schema, and their performance is reported as the micro-average across all events. *N* for events represents the number of sentences in a given set (development, test) for that event, while for slots *N* represents to the number of tagged spans across all sentences in a given set. Sentences occasionally contain more than one event, and as such the number of slots (e.g. SatelliteName) can exceed the number of sentences.

(Perry, 2021) for annotation and calculating interannotator agreement statistics.

Four annotators completed the labeling task, with each sentence being independently labelled by three annotators. After this labelling, the annotators then used the review functionality of LightTag to discuss and resolve disagreements. Individual agreement relative to this final consensus ranged between 80% to 90% precision and 78% to 87% recall across the four annotators.

Summary statistics of the labelled sets are shown in Table 1. The final corpus contains a total of 73.6k tokens across 1,787 sentences, with 15.9k tokens (22%) containing an event slot label.

4. Modelling

4.1. Multi-slot Model

Recently a number of off-the-shelf methods for state-of-the-art triple extraction have emerged (Han et al., 2019), but few of these systems work in multi-slot scenarios. Several groups (Zhang et al., 2020; Liu et al., 2019) demonstrate multi-slot extraction using pre-trained embedding models such as BERT (Devlin et al., 2019), but their systems are not publicly available. We recreate a similar system by framing the multi-slot extraction problem as a sequence labeling task where entity spans are labeled with a given slot name (e.g. SATELLITENAME, LAUNCHVEHICLE, etc.), and adapt an off-the-shelf BERT-based sequence labeling system[7] to this task, which achieves near state-of-the-art performance on entity labeling tasks (Smith et al., 2020).

[7] `https://github.com/kamalkraj/BERT-NER`

4.2. Results

Each of the three events was trained and evaluated separately. We made use of the `BERT-Base-cased` (110M parameter) model. Performance was tuned on the development set, where we observed that performance peaked at 60 training epochs. We report performance using the standard definitions of precision, recall, and F1 (Manning and Schutze, 1999, inter alia).

Overall extraction performance for each SSA event is shown in Table 2. Overall performance per slot ranges from 47 to 91 F1, with higher-frequency slots that tend to follow regular patterns tending to perform better than low-frequency categories with less regular patterns. SATELLITENAME, LAUNCHVEHICLE, and LAUNCHSITE were generally able to achieve fair to excellent performance (peaking near 90 F1 for the *Launch* event), while slots with comparatively less training data and higher variation in their presentation in sentences (e.g. TARGETORBIT and FAILURETYPE) achieved modest performance approaching 74 and 56 F1 on the test set, respectively. FAILURETYPE is the lowest-performing slot, highlighting the challenge of identifying spans of text that describe failures when they may present as either relatively frequent high-level causes (e.g. *"launch failure"* or *"power failure"*), or more-specific descriptions less-frequently observed during training (e.g. *"fuel leak"*, or *"problems with antenna"*) that more precisely convey critical issues.

Prop.	Error
43%	Mention not relevant to event
21%	Span errors (too long or short)
8%	Inferred label is possibly relevant
7%	Gold label is incorrect
5%	Abbreviations mistaken as satellite name

Table 3: Common categories of errors on a cross-section of 100 randomly selected errors on the test set. Proportions do not sum to 100% as less frequent error classes are omitted.

4.3. Error Analysis

To better understand the challenges this dataset poses to extraction systems, we analyzed 100 randomly-selected extraction errors on the test set and identified five major classes of error, outlined in Table 3, and described below:

Mention Not Relevant (43%): Sentences containing space events are frequently long and contain historical or other contextual information that serve as distractors for the events being extracted. For example, in the FAILURE sentence *"AMC-14 was delayed to February due to the failure of a Proton [LaunchVehicle] rocket in September [Date]"*, the model populates the SATELLITENAME slot with AMC-14 even though it is not involved in the failure event.

Span Errors (21%): Span errors are when the model captures part of a mention, but does not overlap completely with the gold information – typically from ending too early, starting too late, or missing tags in the middle of a span (for example, missing *"of"* in *"United States of America"*.

Possibly Relevant (8%): While not identical to gold spans, in 8% of cases, the spans chosen by the model could also be considered valid by manual judgement. For example, in one error, the model broke the gold TARGETORBIT span *"highly elliptical and highly inclined orbit"* into two separate mentions *"highly elliptical"* and *"highly inclined orbit"*, that are both valid.

Errors in gold labels (7%): In 7% of errors, the gold annotation had technical errors. For example, sentences automatically extracted from news articles occasionally contain extraneous out-of-sentence information (such as part of the headline, or captions from photographs) that might contain the same entities as those in the event, and our annotation protocol specifies these extraneous spans not to be annotated – but sometimes the borders are difficult to determine. With the overall micro-precision of the model at 80% across test set events, and 7% of model errors due to annotation errors, we estimate the total accuracy of this annotation after the review procedure to be approximately 98%.[8]

Abbreviations as Satellites (5%): Abbreviations in text (e.g. *"AFP"*, a press organization) are occasionally mistaken as SATELLITENAME slots due to the model learning that satellite names are frequently capitalized.

5. Conclusion

We present the first Space Situational Awareness event corpus and extraction system that can monitor news articles and extract three high-impact events: spacecraft launches, failures, and decommissionings. The corpus contains 1,787 labelled sentences with 15.9k manually labelled tokens, drawn from a corpus of nearly 48.5k news articles spanning all 4,157 known satellites active between 2009 and 2020. Our analysis shows that baseline model performance (F1) ranges between 53 and 91 per event slot, highlighting the challenges associated with this low-resource domain. The corpus and extraction system are open source, available at: `https://github.com/cognitiveailab/ssa-corpus`.

6. Bibliographical References

Cai, H., Hussein, I., and Jah, M. (2020). Possibilistic admissible region using outer probability measure theory. *Acta Astronautica*, 177:246–257.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Esteva, M., Xu, W., Simone, N., Gupta, A., and Jah, M. (2020). Modeling data curation to scientific inquiry: A case study for multimodal data integration. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 235–242.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., and Sun, M. (2019). OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pages 169–174.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450.

Le May, S., Carter, B., Gehly, S., Flegel, S., and Jah, M. (2020). Representing and querying space object

[8]Approximately 20% of tags predicted by the model were errorful on the test set. With a manually estimated 7% error rate in gold labels on these errors, and assuming correctly predicted gold labels are correct labels, we estimate the overall error rate in annotating gold labels to be approximately 2%.

registration data using graph databases. *Acta Astronautica*.

Liu, X., Huang, H.-Y., and Zhang, Y. (2019). Open domain event extraction using neural latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Perry, T. (2021). LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Smith, H., Zhang, Z., Culnan, J., and Jansen, P. (2020). Scienceexamcer: A high-density fine-grained science-domain corpus for common entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4529–4546.

Valenzuela-Escárcega, M. A., Hahn-Powell, G., and Surdeanu, M. (2016). Odin's runes: A rule language for information extraction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 322–329.

Valenzuela-Escárcega, M. A., Hahn-Powell, G., and Bell, D. (2020). Odinson: A fast rule-based information extraction framework. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2183–2191.

Walls, R., Gaylor, D., Reddy, V., Furfaro, R., and Jah, M. (2016). Assessing the iadc space debris mitigation guidelines: A case for ontology-based data management. In *Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS)*.

Wang, X., Feng, A., Golshan, B., Halevy, A., Mihaila, G., Oiwa, H., and Tan, W.-C. (2018). Scalable semantic querying of text. *Proceedings of the VLDB Endowment*, 11(9):961–974.

Zhang, R., Yang, W., Lin, L., Tu, Z., Xie, Y., Fu, Z., Xie, Y., Tan, L., Xiong, K., and Lin, J. (2020). Rapid adaptation of bert for information extraction on domain-specific business documents. *arXiv preprint arXiv:2002.01861*.