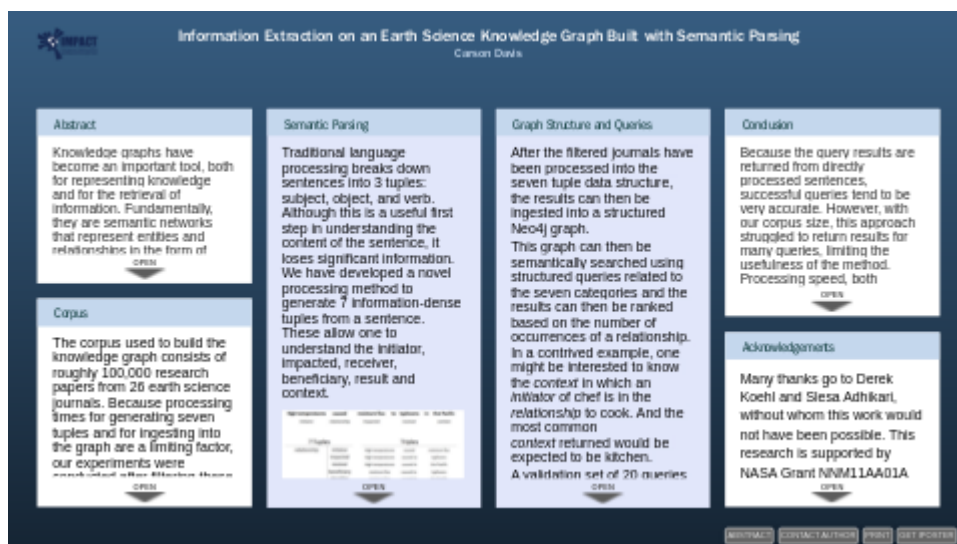


Information Extraction on an Earth Science Knowledge Graph Built with Semantic Parsing



Carson Davis (1), Derek Koehl (1), Slesha Adhikari (1), Muthukumaran Ramamsubramanian (1), Iksha Gurung (1), Rahul Ramachandran (2)

1 - University of Alabama at Huntsville, 2 - National Aeronautics and Space Administration

PRESENTED AT:



ABSTRACT

Knowledge graphs are an important tool, both for representing knowledge and for retrieving information. Fundamentally, they are semantic networks that represent entities and relationships in the form of nodes and edges. A large corpus of natural language text can be broken down into discrete entities and relationships to form a useful knowledge graph. Existing research breaks down text into a subject, object, and verb relationship triple. Although this is a useful first step, it loses much of the original contextual information encoded within the text. Our process uses a novel 7-tuple approach, in which elements of sentences are programmatically parsed into seven categories: initiator, impacted, receiver, beneficiary, result, and context. In this presentation, we show a knowledge graph built using this 7-tuple processing of an Earth science corpus. We explain the techniques used to create the graph and analyze its information retrieval capability while assessing the accuracy and limitations of the results.

CORPUS

The corpus used to build the knowledge graph consists of roughly 100,000 research papers from 26 Earth science journals. Because processing times for generating 7-tuples and for ingesting into the graph are a limiting factor, our experiments were conducted after filtering these journals to only include papers which contained the keywords smoke, biomass burning, and particulate matter, which is an area of expertise for one of team members. This resulted in a refined corpus of roughly 4,000 documents which were then processed and ingested.

Journal Name	Date Published																	
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Atmospheric Science Letters		5	21	34	27	22	42	39	59	64	34	80	69	67	69	65	67	18
Earth and Space Science													1	24	28	25	42	88
Earth's Future												13	26	24	52	51	80	56
Eos, Transactions American Geophysical Union										46	37			1				
Geochemistry, Geophysics, Geosystems	88	207	266	173	191	143	228	174	164	203	192	247	196	154	185	167	64	29
GeoHealth																22	16	14
Geophysical Research Letters	1,151	1,436	1,518	1,688	1,700	1,533	1,559	1,290	1,005	1,234	1,058	1,154	1,200	1,388	1,491	1,390	1,507	1,136
Global Biogeochemical Cycles	132	136	124	120	75	66	94	66	83	64	96	76	72	76	75	74	23	21
Journal of Advances in Modeling Earth Systems								6	3	9	28	35	49	88	66	111	125	134
Journal of Geophysical Research						56					20							
Journal of Geophysical Research: Atmospheres	873	1,238	746	706	784	965	972	737	954	948	782	969	811	748	841	758	784	518
Journal of Geophysical Research: Biogeosciences				23	79	140	191	112	146	185	130	92	109	118	138	132	45	24
Journal of Geophysical Research: Earth Surface		12	47		91	84	145	130	113	134	132	141	117	92	93	59	44	30
Journal of Geophysical Research: Oceans	253	497	375	314	317	341	434	323	382	511	418	338	341	330	325	388	402	284
Journal of Geophysical Research: Planets	137	219	125	126	192	150	195	130	147	171	172	107	89	78	75	103	142	92
Journal of Geophysical Research: Solid Earth	345	603	365	319	377	435	436	376	508	450	378	297	316	307	314	388	126	96
Journal of Geophysical Research: Space Physics	439	561	498	525	475	447	589	533	756	814	694	490	503	541	592	542	466	446
Meteorological Applications								7	47	69	76	76	46	61	67	76	1	2
Paleoceanography	65	109	98	82	67	61	85	47	64	59	55	45	59	76	56	50		
Paleoceanography and Paleoclimatology																	24	26
Quarterly Journal of the Royal Meteorological Society									6	105		203	168	197	178			
Radio Science	108	132	146	114	94	122	91	108	59	136		79	52	69	70	109	93	94
Reviews of Geophysics	9	23	12	11	8	14						12	12	14	23	16	22	17
Space Weather		16	57	53	48	47	44	48	48	55	63	65	53	67	68	88	125	58
Tectonics	55	85	79	88	58	73	66	60	73	67	78	58	74	83	99	119	41	18
Water Resources Research	307	396	319	337	328	364	414	356	350	403	408	447	412	365	403	397	453	346

Distribution of articles across time and journals.

SEMANTIC PARSING

Traditional language processing breaks down sentences into 3-tuples: subject, object, and verb. Although this is a useful first step in understanding the content of the sentence, it loses significant information. We have developed a novel processing method to generate seven information-dense tuples from a sentence. These allow one to understand the initiator, impacted, receiver, beneficiary, result, and context.

High temperatures		caused	moisture flux	to	typhoons	in	the Pacific
initiator		relationship	impacted		receiver		context

7 Tuples		Triples		
relationship	initiator	High temperatures	caused	moisture flux
	impacted	High temperatures	caused to	typhoons
	receiver	High temperatures	caused in	the Pacific
	beneficiary	moisture flux	caused to	typhoons
	identifier	moisture flux	caused in	the Pacific
	context	typhoons	caused to in	the Pacific

This chart shows how an earth science sentence would be broken into each of the seven tuples. By using the relationship as a connector, each of the six remaining, populated tuple positions can be joined in a meaningful triple, which can then be graphed and queried.

GRAPH STRUCTURE AND QUERIES

After the filtered journals have been processed into the 7-tuple data structure, the results can then be ingested into a structured Neo4j graph.

This graph can then be semantically searched using queries that are structured around the seven categories and the results are then ranked based on the number of occurrences of a relationship. In a contrived example, one might be interested to know the context in which an initiator of chef is in the relationship to cook. And the most common context returned would be expected to be kitchen.

A validation set of 20 queries was created by first creating plain language questions related to the initial filter keywords of smoke, biomass burning, and particulate matter. These plain language questions were then manually translated into structured graph queries. Below you can see two successful example queries.

Below you can see two successful example queries, sourced from a validation set of 20 queries.

Query Process

How does smoke impact solar radiation?

- Initiator
 - smoke
- Impacted
 - solar radiation
- search for Relationships

In what context does biomass burning occur?

- Initiator
 - biomass burning
- Relationship
 - occur
- search for Context

Query Results

How does smoke impact solar radiation?

- absorbs - 4
- attenuates - 2
 - by scattering
- reduced - 1
 - by w m-2

In what context does biomass burning occur?

- in northern australia - 2
- in australian tropics - 1
- in tropical regions - 1

The results of both structured queries represent accurate answers to the original plain-language questions.

CONCLUSION

Because the query results are returned from directly processed sentences, successful queries tend to be very accurate. However, with our corpus size, this approach struggled to return results for many queries, limiting the usefulness of the method. Processing speed, both through the 7-tuple pipeline and ingesting into the interconnected graph, is very time consuming and presents a second limitation. If this approach could be applied to a much larger corpus and processing could be sped up, it could present an effective query tool.

ACKNOWLEDGEMENTS

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program.

Notice!

Your iPoster has now been unpublished and will not be displayed on the iPoster Gallery.

You need to publish it again if you want to be displayed.

ABSTRACT

Knowledge graphs have become an important tool, both for representing knowledge and for the retrieval of information. Fundamentally, they are semantic networks that represent entities and relationships in the form of nodes and edges. A large corpus of natural language text can be broken down into various entities and relationships to form a useful knowledge graph. Existing research breaks down text into a subject, object, and verb relationship triple. Although this is a useful first step, it loses much of the original contextual information encoded within the text. Our process uses a novel 7-tuple approach, in which elements of sentences are programmatically parsed into seven categories: initiator, impacted, receiver, beneficiary, result, and context. In this presentation, we show a knowledge graph built using this 7-tuple processing of an Earth science corpus. We explain the techniques used to create the graph and analyse its information retrieval capability while assessing the accuracy and limitations of the results.