[FC-03-032] Semantic Information Elicitation

Abstract

The past few decades have been characterized by an exponential growth of digital information resources. A considerable amount of this information is semi-structured, such as XML files and metadata records and unstructured, such as scientific reports, news articles, and historical archives. These resources include a wealth of latent knowledge in a form mainly intended for human use. Semantic information elicitation refers to a set of related processes: semantic information extraction, linking, and annotation that aim to make this knowledge explicit to help computer systems make sense of the content and support ontology construction, information organization, and knowledge discovery.

In the context of GIScience research, semantic information extraction aims at processing unstructured and semi-structured resources and identifying specific types of information: places, events, topics, geospatial concepts, and relations. These may be further linked to ontologies and knowledge bases to enrich the original unstructured content with well-defined meaning, provide access to information not explicit in the original sources, and support semantic annotation and search. Semantic analysis and visualization techniques are further employed to explore aspects latent in these sources such as the historical evolution of cities, the progression of phenomena and events and people's perception of places and landscapes.

Keywords: ontology, semantics, semi-structured text, structured text, unstructured text

Author & citation

Kokla, M. (2021). Semantic Information Elicitation. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2021 Edition), John P. Wilson (ed.).
DOI: 10.22224/gistbok/2021.2.10.

Explanation

1. Definitions

Structured data: highly organized data, such as spreadsheets and databases with well-defined schema.

Semi-structured data: data with some degree of organization and structure, such as XML files.

Unstructured data: data lacking in organization and structure, such as text documents, audio, video, and images.

Semantic information extraction: the process of identifying mentions of entities, concepts, topics, and relations in a given unstructured or semi-structured source.

Semantic annotation: the process of tagging semi-structured or unstructured input sources with entities, concepts, topics, and relations, typically based on an existing ontology or knowledge base.

Ontology learning: the process of creating a new ontology or enriching an existing ontology with additional concepts and relations from a collection of documents describing a domain.

Ontology population: the process of enriching an existing ontology with new instances of concepts extracted from a collection of documents.

2. Semantic Information Elicitation

Semantic information elicitation may refer to a set of processes that draw out latent knowledge from unstructured or semi-structured content (Kokla & Guilbert, 2020). Such content includes a wealth of latent spatial knowledge in terms of places, events, spatial concepts, and relations. Semantic information extraction, linking, and annotation aim to elicit a structured representation of the semantic content of such resources which may be further linked to ontologies and knowledge bases to enrich the original unstructured content with well-defined meaning and provide access to information not explicit in the original sources.

The rising requirement for ontologies to provide machine readable annotations for the support of the Semantic Web has also led to the development of approaches for ontology construction from text (Al-Aswadi et al., 2020). Ontology learning is the process of constructing a new ontology or enriching an existing ontology with concepts and relations, whereas ontology population is the process of adding new instances of concepts to an existing ontology.

Semantic structures may also be elicited from domain experts using crowdsourcing approaches, like the bottom-up development, editing, and validation of GIS&T BoK concepts through a collaborative visual wiki environment (Ahearn et al., 2013).

The result of semantic information elicitation processes may take various forms: semantic tags, topics, keywords, entities, and ontology concepts and relations. These outputs may be further exploited in different ways to support content-based exploration, semantic search, data-driven analysis and visualization.

3. Semantic Information Extraction and Semantic Annotation

Semantic information extraction refers to a range of tasks used to identify specific types of information from semi-structured and unstructured sources. Three main types of elements are extracted: entities, concepts, and relations using various underlying sub-tasks such as Named Entity Recognition and Disambiguation, Concept Extraction, Relation Extraction, etc.

(Figure 1).

A subfield of information extraction, called ontology-based information extraction (OBIE), uses the formal specification of domain knowledge provided by an ontology to guide the extraction of pre-defined concepts, properties, and relations (Karkaletsis et al., 2011). OBIE includes approaches that: (a) use an ontology as a guide to acquire knowledge from text; and (b) ontology learning and population approaches that seek to build or enrich an ontology by processing texts (Wimalasuriya & Dou, 2010).

OBIE is closely related to semantic annotation (or tagging), a process which focuses on enriching documents and other unstructured data sources with entities, concepts, topics, and relations. Semantic annotation involves the identification of these elements in the input sources and their linking to relevant unique elements in an ontology or knowledge base (KB) (Liao & Zhao, 2019). Semantic annotation reinforces the integration, reuse, search, and discovery of information and is therefore crucial for the development of the Semantic Web.

Semantic annotation of texts consists of subtasks including preprocessing, named entity recognition and linking, concept extraction and linking, and relation detection and linking. Several of these subtasks are common to both information extraction and semantic annotation (Figure 1) and are outlined in the next sections.

Besides texts, other types of resources may be semantically annotated, such as photos (Ennis et al., 2015), videos (Nixon et al., 2013), and maps (Hu et al., 2015). In that case, ontologies and other knowledge sources are used to generate semantic annotations of places, concepts, and events in these resources and thus support the semantic interpretation of the original content that goes beyond object detection and classification.

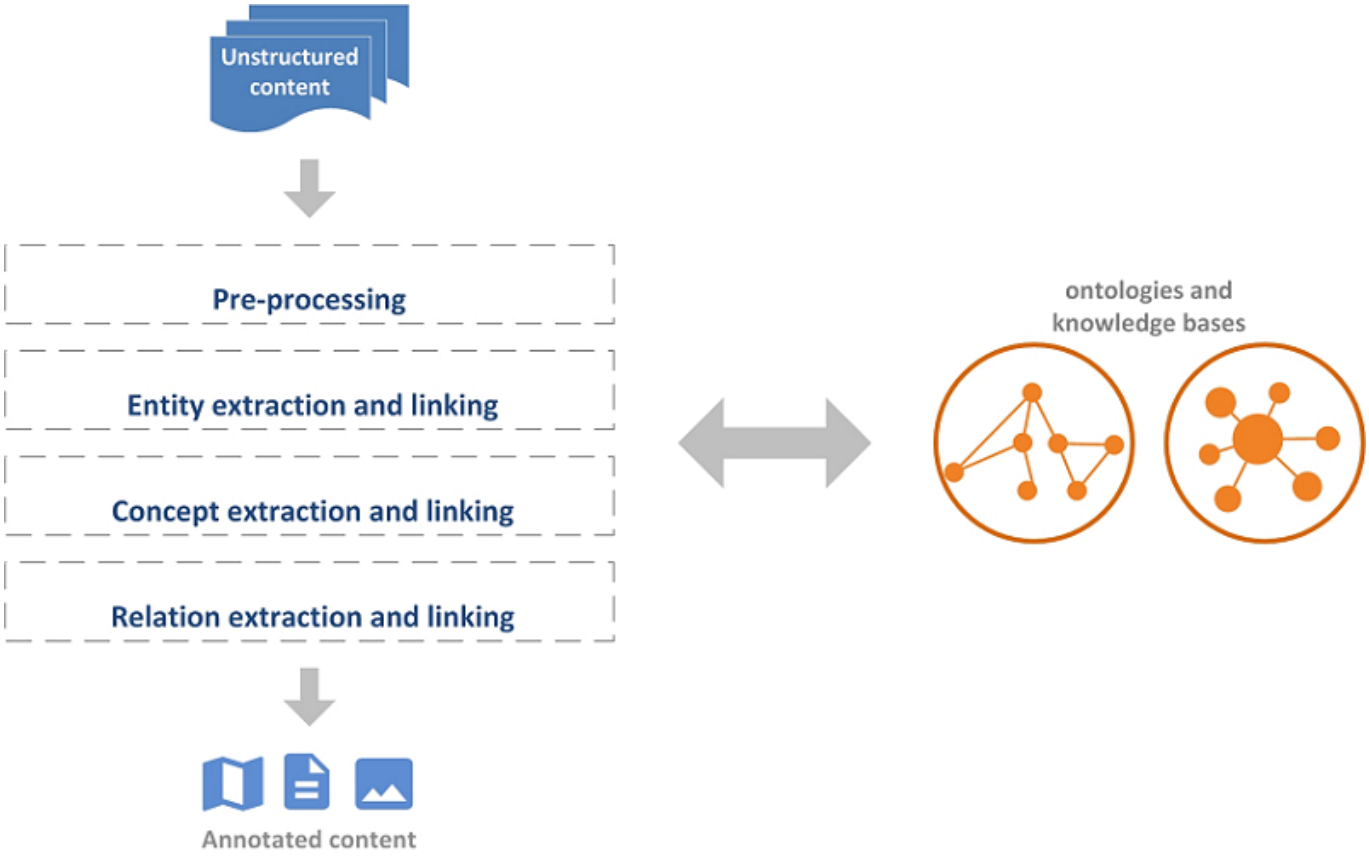

Figure 1. Semantic information extraction and semantic annotation subtasks. Source: author.

3.1 Entity Extraction and Linking

Named entity recognition (NER) is an information extraction task used to detect named entity mentions in text and classify them in pre-defined categories, such as persons, locations, organizations, dates, etc. Named entities refer to specific real-world objects or instances and are typically denoted by proper nouns or temporal expressions. NER is commonly based on gazetteers, i.e., lists that include words or phrases representing individual instances of a specific category (e.g., person, location, time, and organization).

In many cases however, the symbols (words) denoting named entities are not unique, for example, ‘the Atlantic’ usually refers to the ocean, but in different contexts it may refer to several places, a magazine, a period of palaeoclimatology and various other entities. In such cases, where the same entity mention refers to different instances, Named Entity Disambiguation (NED) is used to identify a unique semantic interpretation for the entity mention.

Entity Linking (EL) is the process of mapping mentions of entities from text to unique entities in a reference KB, such as Wikipedia, DBpedia, or YAGO and is often used to support semantic annotation.

Place names are crucial for indirect geographic referencing in many geospatial applications (Hill, 2000). Geoparsing is the process used to extract and disambiguate place names (also known as toponyms) from natural language texts and connect these to a unique location on

Earth using geographic coordinates (Gritta et al., 2018).

Besides place names, some applications require the extraction of other entities such as temporal expressions for the identification and analysis of events, phenomena, or movements. Figure 2 shows a simple example of entity extraction and linking. Various mentions of locations (Sumatra, Indonesia, Sri Lanka, etc.), a date (December 26, 2004) and an instance of an earthquake (2004 Indian Ocean earthquake) may be identified and classified into relevant categories (location, date, and earthquake). Then, these entities (in the example Indonesia, Sumatra, and 2004 Indian Ocean earthquake) are associated with corresponding entities in BabelNet and GeoNames to further enrich their semantic description with information (e.g., continent, country, population, geographic coordinates, etc.).

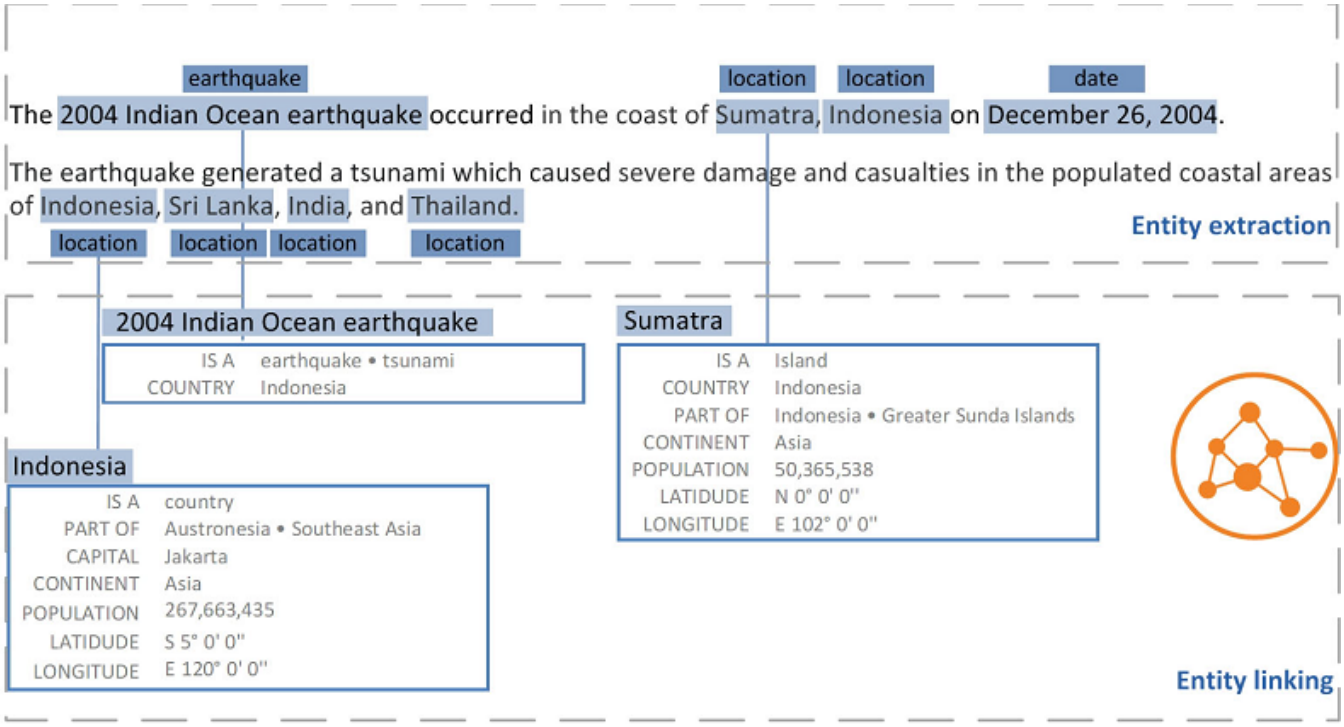

Figure 2. Entity extraction and linking. Source: author.

3.2 Concept Extraction and Linking

Concept extraction refers to the detection of words or phrases that express semantic classes of elements. Concept extraction involves three related processes: terminology extraction, keyword/key phrase extraction, and topic modeling (Martinez-Rodriguez et al., 2018). Terminology or term extraction detects the central terms / concepts of a given domain described by a text collection. For example, a text collection on climate change will probably include domain concepts, such as greenhouse gases, global warming, and ocean acidification. Keyword / key phrase extraction on the other hand consists in extracting words or phrases that describe the subject or domain of a given document.

Terminology and key phrase extraction begin by selecting candidate domain terms. This step is usually performed by applying pre-processing tasks such as tokenization to split text

into words, phrases, symbols, or other elements called tokens, part-of-speech (POS) tagging to assign a particular part of speech to a given word or phrase, and lemmatization to identify the base or dictionary form of a word (lemma). Then, contextual information within a window of predefined length around a term is considered to determine whether the term may represent a domain concept. Alternatively, a set of rules that specify regular expressions and lexico-syntactic patterns are used for extracting predefined domain concepts. Candidate terms are subsequently filtered to select those that best describe a domain or the subject of a text using linguistic or statistical methods. The concepts extracted may be further linked to a reference ontology or KB.

Figure 3 shows an example of extracting concepts such as earthquake, tsunami, and coastal areas. Then, linking these concepts to relevant concepts in a reference ontology or KB such as BabelNet results in enriching the knowledge about the concepts with further semantic information (e.g., the concept ‘earthquake’ may be enriched with a definition, instances, and properties such as measurement scale).

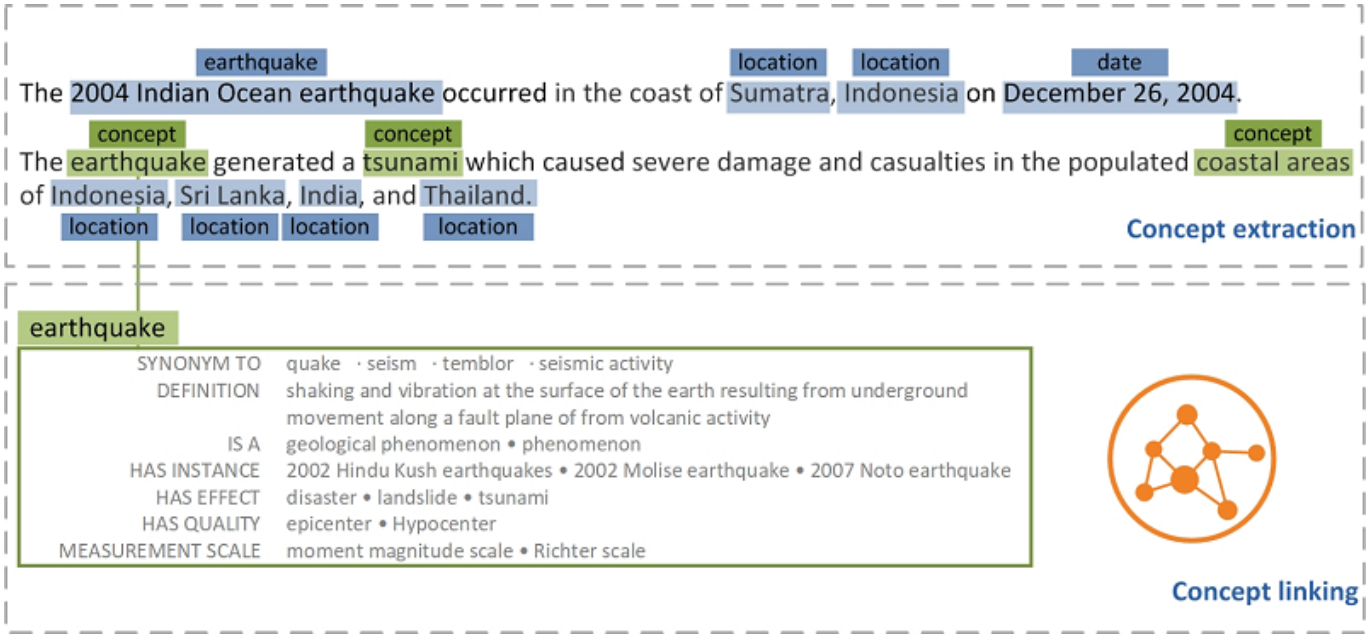

Figure 3. Concept extraction and linking. Source: author.

Topic modeling may be considered as a closely related task that aims at discovering latent semantic topics and associations from large text collections. Topic modeling such as Latent Dirichlet Allocation (LDA) are text mining techniques employed to identify terms that frequently co-occur in a given text collection and form clusters that represent abstract topics (Blei, 2012). Topic labeling is then used to identify words or phrases that describe the abstract topic. For example, Figure 4 shows three topics that characterize a set of articles from the GIS&T BoK.

Figure 4. Three abstract topics and their top 20 most frequent words learned by the LDA model using MALLET (McCallum, 2002) for a subset of GIS&T BoK: (a) programming languages for GIS applications, (b) GIS&T and society and (c) cartography and visualization. Source: author.

3.3 Relation Extraction and Linking

Terminology and keyphrase extraction may also include the detection of semantic relations between domain concepts such as hierarchical (hypernymy) and/ or similarity (synonymy) relations. Relations between entities can also be extracted and modeled. In that case, relation extraction refers to the identification and classification of various, mostly binary, instances of relations between named entities (e.g., located-at, caused-by, etc.). In case the relations extracted from text are further linked to a reference KB, the process is called Relation Linking (EL).

The process followed by relation extraction and linking (REL) may vary (Martinez-Rodriguez et al., 2018). REL approaches involve the extraction and linking of named entities and concepts which often precede the extraction of relations, and parsing of relations from text. Existing relations derived from a reference KB may be used to identify general patterns and guide the extraction of novel relations.

Documents include wealth of spatial relations which are expressed both quantitatively and qualitatively using spatial expressions such as at, near, next to, to the right of, south of, etc. Spatial relations are usually modelled as triples that include an object to be located, a reference object, which are either a place name or a geographic feature type, and a spatial relation between them.

Figure 5 shows a spatial relation (occurred-in) between the entities ‘2004 Indian Ocean earthquake’ and ‘Sumatra’ and a cause-effect relation between the concepts ‘earthquake’ and ‘tsunami’. Linking the extracted entities and concepts to BabelNet allows the identification of an IS-A relation between the entity ‘2004 Indian Ocean earthquake’ and the concepts ‘earthquake’ and ‘tsunami’ (Figure 5), as well as the identification of semantic relations (such as is-a, has effect, or has quality) for the concept ‘earthquake’ (Figure 3).

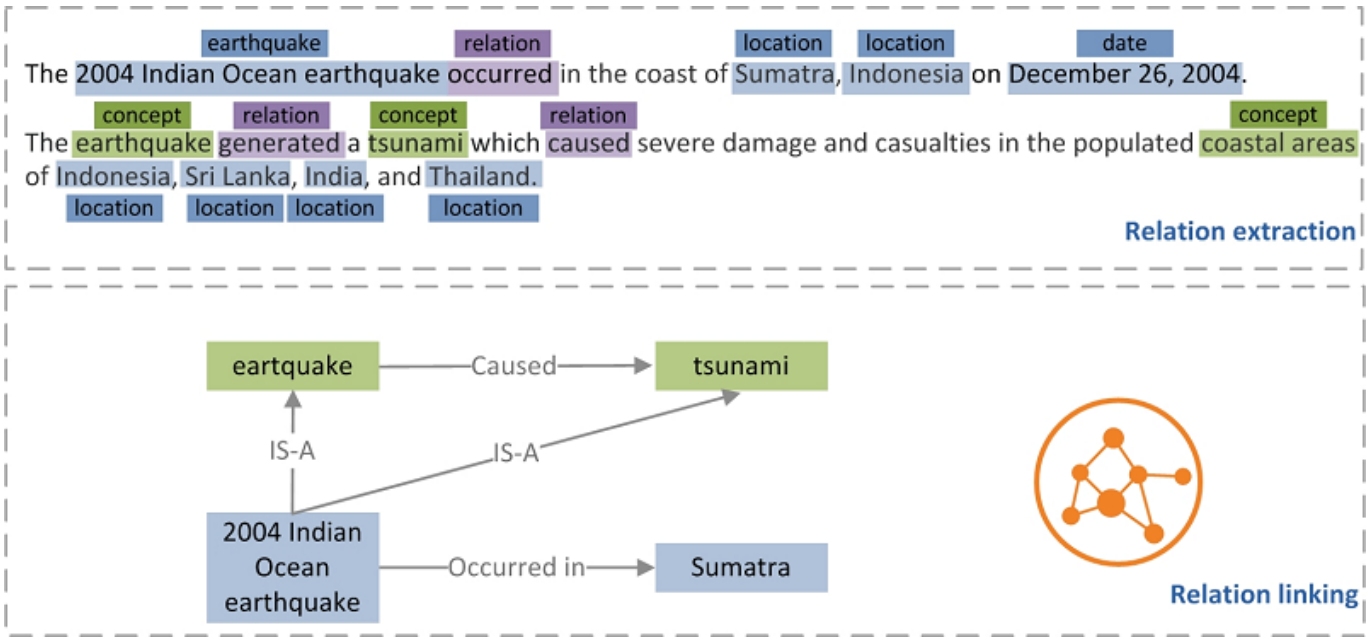

Figure 5. Relation extraction and linking. Source: author.

Complex information referring to events, such as natural hazards and movements can also be extracted from texts through the combined extraction of spatial and temporal information involving place names, temporal expressions, and spatio-temporal relations.

3.4 Sentiment Analysis and Emotion Detection

Besides named entities, concepts, relations and topics, subjective elements, such as sentiments, emotions, and opinions, can be detected from the information available on the web. In general, sentiment analysis or opinion mining aims to analyze people's emotions, opinions, and attitudes concerning specific topics, persons, products, and organizations (Serrano-Guerrero et al., 2015). Sentiment analysis may be performed either at the level of the whole text or at the level of a sentence. Sentiments are analyzed according to their polarity into three categories (positive, negative and neutral) or more resulting in more detailed polarity ratings.

Emotion detection (ED) (Acheampong et al., 2020) is a subfield of sentiment analysis that focuses on the extraction and analysis of emotions, such as anger, disgust, fear, joy, sadness, and surprise. For example, Figure 6 shows an example of sentiment analysis and emotion detection. In this case, sadness, an emotion with negative polarity is detected. Emotions are used to provide fine-grained insights into people's personal experiences and their interactions with their surroundings. In this context, people's emotions associated with places or other geographic features or concepts are detected and modeled in order to explore emotional patterns in large text collections and delve into the subjective representation of these features, based on human experience (Ballatore & Adams, 2015).

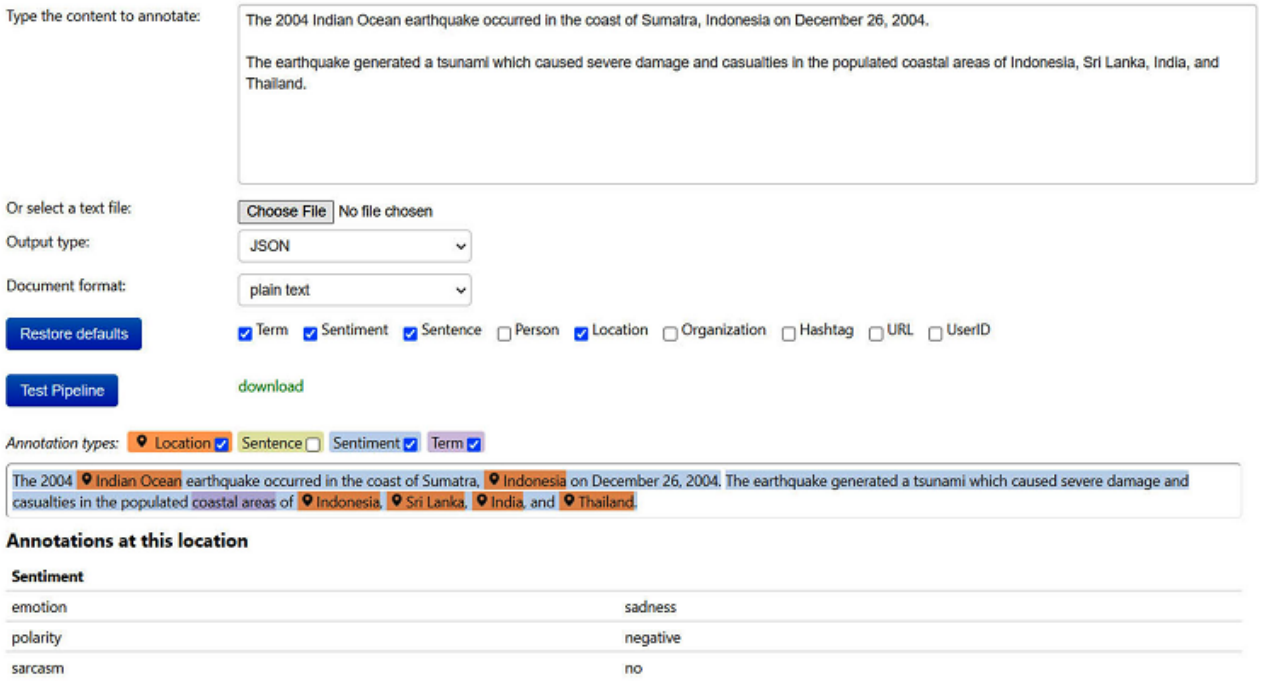

Figure 6. Example of semantic annotation using the DecarboNet Environmental Annotator. The annotator extracts locations and terms related to climate change. It also identifies sentiments which are classified into positive, negative or neutral and detects more fine-grained emotions such as fear, anger, joy, etc. Terms related to climate change are also linked (where appropriate) to relevant Linked Open Data ontologies. Source: DecarboNet Environmental Annotator.

4. Semantic Analysis and Visualization

The semantic information extracted from semi-structured and unstructured sources may be further explored for the semantic analysis and visualization of these sources. A variety of input sources such as historical texts, travel blogs, and social media is used to identify various elements relevant to the semantic description of geographic features, such as topics and thematic characteristics associated with places and regions and the variations of the semantic descriptions across space and time. For example, the extraction of place names combined with temporal information and abstract topics such as education, economy, politics, art, etc. may be used to analyze the importance of a place and its spatio-temporal evolution, or the economic, political, and cultural relations or similarities between places, given the identified topics (Salvini & Fabrikant, 2016).

The significant amount of natural language texts on the web, such as social media and blogs, also provide an alternative source to study people's perceptions of places, landscapes and other geospatial features following a bottom-up data-driven approach. These analyses, despite their limitations in terms of representativeness, may be used complementarily to conventional human-participants studies (Gao et al., 2017) to reveal social, cultural, and perceptual aspects, emotions, and sense of place from user-generated content.

Cartographic techniques such as density surfaces or spatialization techniques (Skupin &

Fabrikant, 2007) such as networks and self-organizing maps may be further employed for the visualization, exploration, and interpretation of large complex unstructured data. Figure 7 shows the Geospatial Technology Explorer, an interactive map for the visualization and exploration of domain concepts, created using a combination of natural language processing, machine learning, and visualization from over 100,000 GIS&T domain documents. The semantic landscape represents domain concepts at different levels of detail and focus; mountains correspond to highly focused domain concepts, whereas valleys to more general mixed concepts.

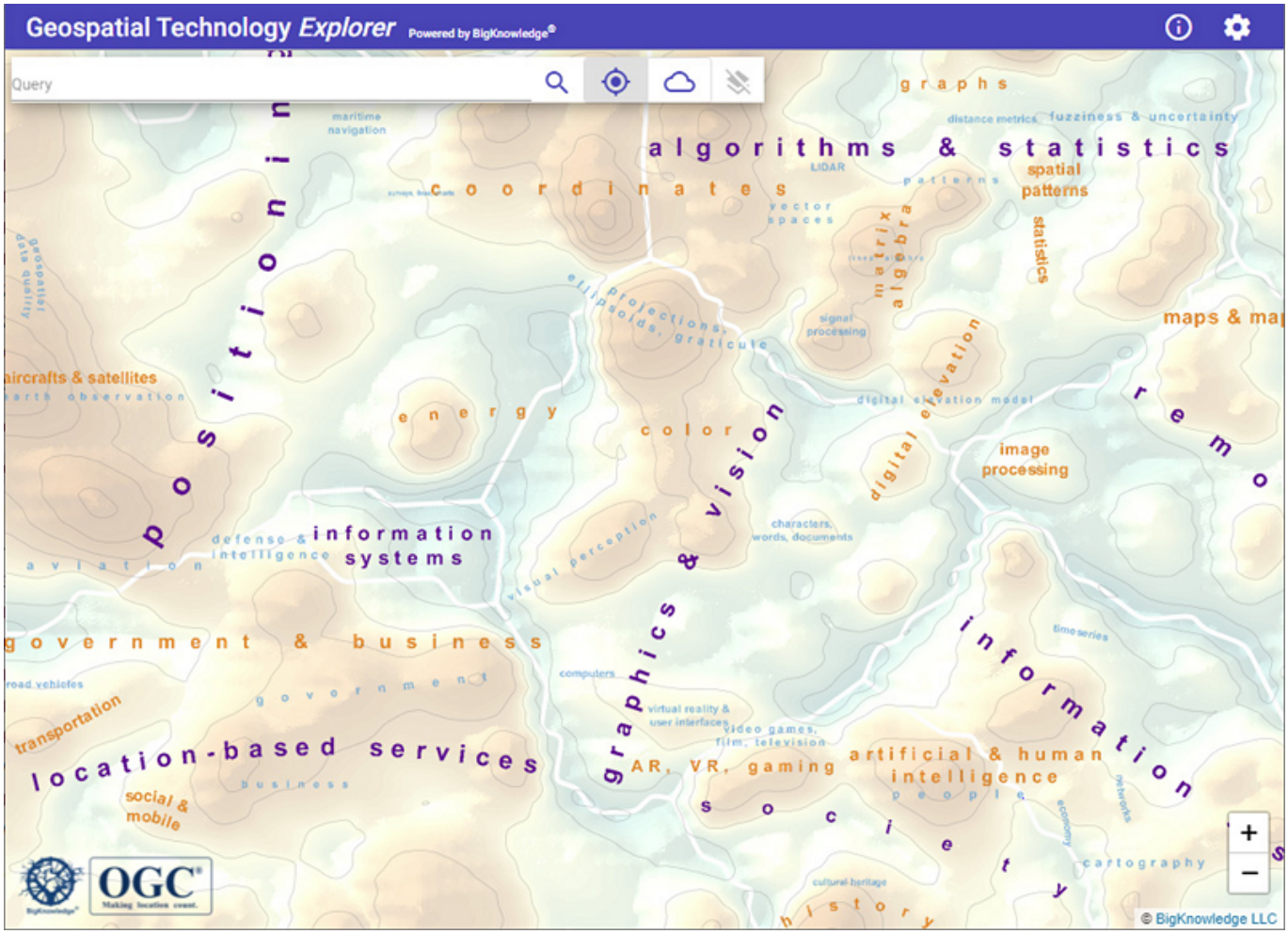

Figure 7. A base map created from over 100,000 GIS&T domain documents and integrated into the Geospatial Technology Explorer, from BigKnowledge LLC and OGC. Source: Copyright BigKnowledge LLC. Used with permission.

References

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), e12189.

Ahearn, S. C., Icke, I., Datta, R., DeMers, M. N., Plewe, B., & Skupin, A. (2013). Re-engineering the GIS&T Body of Knowledge. International Journal of Geographical Information Science, 27(11), 2227–2245.

Al-Aswadi, F. N., Chan, H. Y., & Gan, K. H. (2020). Automatic ontology construction from

text: A review from shallow to deep learning trend. Artificial Intelligence Review, 53(6), 3901–3928.

Ballatore, A., & Adams, B. (2015). Extracting Place Emotions from Travel Blogs. In F. Bacao, M. Y. Santos, & M. Painho (Eds.), AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities, Lecture Notes in Geoinformation and Cartography (pp. 1– 5). Springer.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Gao, S., Janowicz, K., Montello, D.R., Hu, Y., Yang, J-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., and Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science, 31:6, 1245-1271.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? Language Resources and Evaluation, 52(2), 603–623.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries (pp. 280–290). Springer.

Karkaletsis, V., Fragkou, P., Petasis, G., & Iosif, E. (2011). Ontology Based Information Extraction from Text. In G. Paliouras, C. D. Spyropoulos, & G. Tsatsaronis (Eds.), Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap (pp. 89–109). Springer.

Kokla, M. & Guilbert, E. (2020). A Review of Geospatial Semantic Information Modeling and Elicitation Approaches. ISPRS International Journal of Geo-Information, 9(3), 146.

Liao, X., & Zhao, Z. (2019). Unsupervised Approaches for Textual Semantic Annotation, A Survey. ACM Computing Surveys, 52(4), 1–45.

Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2018). Information extraction meets the Semantic Web: A survey. Semantic Web, 1–81.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit.

Salvini, M. M., & Fabrikant, S. I. (2016). Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design, 43(1), 228–248.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18–38.

Skupin, A., & Fabrikant, S. I. (2007). Spatialization. In J. P. Wilson & A. S. Fotheringham (Eds.), The Handbook of Geographic Information Science (pp. 61–79). Blackwell Publishing Ltd.

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36(3), 306–323.