Knowledge will Propel Machine Understanding of Content: Extrapolating from Current Examples

Amit Sheth, Sujan Perera, and Sanjaya Wijeratne

Kno.e.sis Center, Wright State University
Dayton, Ohio, USA
{amit,sujan,sanjaya}@knoesis.org
http://www.knoesis.org

Abstract. Machine Learning has been a big success story during the AI resurgence. One particular stand out success relates to unsupervised learning from a massive amount of data, albeit much of it relates to one modality/type of data at a time. In spite of early assertions of the unreasonable effectiveness of data, there is increasing recognition of utilizing knowledge whenever it is available or can be created purposefully. In this paper, we focus on discussing the indispensable role of knowledge for deeper understanding of complex text and multimodal data in situations where (i) large amounts of training data (labeled/unlabeled) are not available or labour intensive to create, (ii) the objects (particularly text) to be recognized are complex (i.e., beyond simple entity – person/location/organization names), such as implicit entities and highly subjective content, and (iii) applications need to use complementary or related data in multiple modalities/media. What brings us to the cusp of rapid progress is our ability to (a) create knowledge, varying from comprehensive or cross domain to domain or application specific, and (b) carefully exploit the knowledge to further empower or extend the applications of ML/NLP techniques. Using the early results in several diverse situations – both in data types and applications – we seek to foretell unprecedented progress in our ability for deeper understanding and exploitation of multimodal data.

Keywords: Semantic analysis of multimodal data, Knowledge-enabled computing, Machine intelligence, Multimodal exploitation, Understanding complex text, Knowledge-enhanced ML and NLP, Knowledge-driven deep content understanding, Semantic search, Domain specific information retrieval, Ontology, Knowledgebases, Background knowledge, EmojiNet, Emoji Sense Disambiguation, Knowledge-aware search, Enhancing statistical models with knowledge, Implicit Entity Linking

1 Introduction

Recent success in the area of Machine Learning (ML) for Natural Language Processing (NLP) has been largely credited to the availability of enormous training datasets and computing power to train complex computational models [11].

Complex NLP tasks such as statistical machine translation and speech recognition have hugely benefitted from web-scale unlabeled data that is freely available for consumption by learning systems such as deep neural nets. However, many traditional research problems related to NLP such as part-of-speech tagging, and named entity recognition (NER), require labeled or human annotated data, where the creation of such datasets is expensive in terms of the human effort required. In spite of early assertion of the unreasonable effectiveness of data (i.e., data alone is sufficient), there is an increasing recognition of utilizing knowledge to solve complex AI problems. A number of AI experts, including Yoav Shoham [34], Oren Etzioni, and Pedro Domingos [8], have talked about this in recent years.

The value of domain/world knowledge in solving complex problems was also recognized much earlier [38, 6, 19]. These early efforts were centered around understanding the language. Hence, the major focus was towards representing linguistic knowledge. The most popular artifacts of these efforts are FrameNet [25] and WordNet [17], which were developed by realizing the ideas of frame semantics [10] and lexical-semantic relations [7], respectively. Both these resources are used extensively by the NLP research community to understand the semantics of natural language.

The building and utilization of the knowledge bases took a major leap with the advent of the semantic web in early 2000s. For example, it was the key to the first patent on Semantic Web and a commercial semantic search/browsing and personalization engine 15 years ago [30], where knowledge in multiple domains complemented ML techniques for information extraction (NER, semantic annotation) and helping to build semantic applications [27]. Major efforts in the semantic web community have produced large, cross domain (e.g., DBpedia, Yago, Freebase, Google Knowledge Graph) and domain specific (e.g., Gene Ontology, MusicBrainz, UMLS) knowledge bases in recent years, as well as the intelligent applications discussed next.

The value of these knowledge bases has been demonstrated with many applications, including semantic similarity [15], question answering [26], ontology alignment, and word sense disambiguation (WSD) [16], as well as major practical AI services, including Apple's Siri, Google's Semantic Search, and IBM's Watson. For example, Siri relies on knowledge extracted from reputed online resources to answer queries on restaurant searches, movie suggestions, nearby events, etc. In fact, "question answering", which is the core competency of Siri was built by partnering with Semantic Web or Semantic Search service providers who extensively utilize knowledge bases in their applications[1]. The Jeopardy version of IBM Watson uses semi-structured and structured knowledge bases such as DBpedia, Yago, and WordNet to strengthen the evidence and answer sources to fuel its DeepQA architecture [9]. Google Semantic Search is fueled by Google Knowledge Graph[2], which is also used to enrich search results similar to what Taalee/Semagix semantic search engine did 15 years ago [27].

[1] `https://en.wikipedia.org/wiki/Siri`
[2] `http://bit.ly/22xUjZ6`

While knowledge bases are used in an auxiliary manner in the above scenarios, we argue that they have a major role to play in understanding real-world data. The real-world data has greater complexity that has yet to be appreciated and supported by automated systems. This complexity emerges from various dimensions. Human communication has added many constructs to language which help people to communicate effectively. However, current information extraction solutions fall short in processing complex constructs. One such complexity is the ability to express ideas, facts, and opinions in an implicit manner. For example, the sentence *"The patient showed accumulation of fluid in his extremities, but respirations were unlabored and there were no use of accessory muscles"* refers to the clinical conditions "shortness of breath" and "edema", which would be understood by a clinician. However, the sentence does not contain names of those clinical conditions, rather it contains descriptions that imply the two conditions. Current literature on entity extraction has not paid much attention to implicitly stated entities.

Another complexity in real-world data is its multimodal nature. A growing number of real world scenarios involve data coming from different modalities, often complementing each other. There is an increasing availability of physical (including sensor/IoT), cyber, and social data related to events and experiences of human interest [28, 35]. For example, in our personalized digital health application for managing asthma in children[3], we use sensors measuring the patient's physiology (e.g., exhaled nitric oxide) and his immediate surroundings (e.g., carbon monoxide, particulate matter, temperature, humidity), data accessible from the Web for the local area (air quality, pollen, weather), and social data (tweets relevant to asthma, web forum data) [1]. Each of these dimensions provide information that are helpful in proving or disproving the hypothesis provided by medical practice and disease management. Hence, understanding the patient status requires interpreting the observations of each modality and establishing the relationship between them to provide a comprehensive picture. Knowledge bases play a major role in establishing the relationships between multiple observations and transcend multiple abstraction levels [29]. We could know the relationship between asthma, nitric oxide and asthma medications through knowledge bases. Unless we have access to that knowledge or process data from all modalities potentially at different levels of abstractions, our predictions would not be accurate. Hence it is imperative for applications such as personalized digital health monitors to process multimodal data and use the available domain knowledge (in our case, a representation of relevant medical protocol) to arrive at accurate decisions. Emoji sense disambiguation is another example for multimodal data analysis, which is discussed later in the paper.

We argue that careful exploitation of knowledge can greatly enhance the current ability of (big) data processing; it can especially help in dealing with complex situations. At Kno.e.sis, we have recently dealt with situations where:

[3] `http://bit.ly/kAsthma`

1. Large quantities of hand-labeled data required for unsupervised (self-taught) techniques to work well are not available or the annotation effort is significant.
2. The text to be recognized is complex (i.e., beyond simple entity - person/location/organization names) and we need deeper understanding than what traditional information extraction gives, such as complex/compound entities [23], implicit entities [21, 22], and subjectivity (emotions, intention) [12, 36].
3. An application can benefit from multimodal data [1, 2, 4].

These efforts, with the exception of compound entities and emotion identification, have centered around exploiting different kinds of knowledge bases and using semantic techniques to complement or enhance ML, statistical techniques, and NLP. Our ideas are inspired by the human brain's ability to learn and generalize knowledge from a small amount of data (i.e., humans do not need to examine tens of thousands of cat faces to recognize the next cat shown to them), analyze situations by simultaneously and synergistically exploiting multimodal data streams, and understand more complex and nuanced aspects of content especially by knowing (through common-sense knowledge) semantics/identity preserving transformations.

2 Challenges in creating/using knowledge bases

Last decade saw an increasing use of background knowledge in solving diverse problems. They heavily used large, publicly available knowledge bases. While applications such as search, browse, and question answering could use these knowledge bases in their current forms, others like movie recommendation, biomedical knowledge discovery, and clinical data interpretation are challenged by the limitations in current knowledge bases. These limitation are three fold;

1. Messiness of the knowledge bases,
2. Incompleteness and insufficiency of knowledge bases, and
3. Limitations in knowledge representation and reasoning techniques.

Messiness of the knowledge bases: Rapid growth in knowledge bases is occuring both in terms of numbers and sizes. This growth challenges the proper organization of the knowledge bases on the Web, hence users of the knowledge bases increasingly finding it hard to find the relevant knowledge bases or the relevant portion from the large knowledge bases for the domain of interest (e.g., movie, clinical, biomedical). This highlights the need for identifying relevant knowledge bases from collection of knowledge bases such as linked open data cloud and extracting relevant portion of the knowledge from large knowledge bases such as Wikipedia and DBpedia. In order to address this problem, we are working on automatically identifying and indexing the domains of the knowledge bases [14] and exploiting the semantics of the entities and their relationships to

identify the relevant portions of a knowledge base given a domain of interest.

Incompleteness and insufficiency of knowledge bases: The existing knowledge bases can be incomplete with respect to a task at hand. For example, applications like computer assisted coding and clinical document improvement require comprehensive knowledge about a particular domain (e.g., cardiology, oncology). We observe that although the existing medical knowledge bases (e.g., Unified Medical Language System (UMLS)) are rich in taxonomical relationships, they lack non-taxonomical relationships among clinical entities. We developed an algorithm that uses real-world clinical data and existing knowledge to discover more relationships between clinical entities using a human-in-the-loop model [20]. This model is capable of enriching the existing clinical knowledge bases with more relationships. Yet another challenge is creating personalized knowledge bases for specific tasks. For example, in [31], personal knowledge graphs are created based on the content consumed by a user, taking into account the dynamically changing vocabulary, and applied to improve subsequent filtering of relevant content.

Limitations in knowledge representation and reasoning techniques: The scope of what is captured in the knowledge bases is rapidly expanding, requiring better modeling and understanding of concepts well beyond entities and relations, and involves subjectivity (intention, emotions, sentiments), temporal information, and more. This expansion challenges the current knowledge representation solutions which are mainly restricted to triple-based representations. The standard knowledge representation languages developed by Semantic Web community (e.g., RDF, OWL) are limited in their expressivity. It is important to be able to express the context of the knowledge represented with triples, to be able to associate probability value for triple based representations signifying uncertainty of the represented fact, and to be able to reason with them. All these requirements are well-recognized by the community and in recent years, we have seen some promising research in these directions. The singleton-property based representation [18] adds ability to make statements about triples (i.e., to express context of the triple) and probabilistic soft logic [13] adds ability to associate probability value with statements (i.e., triples) and reason over them. It will be really exciting to see applications exploiting such enhanced knowledge representation models that perform 'human-like' reasoning on them.

Next, we will present two new research applications that utilize knowledge bases and multimodal data to address some of the aforementioned challenges identified earlier (i.e., complex nature of the problem, and insufficient manually created knowledge).

Example 1: Implicit entity linking

One of the complexities with data is the ability to express facts, ideas, and opinions in an implicit manner. As humans, we seamlessly use implicit constructs

in our daily conversations and rarely find it difficult to decode the content of the messages. Consider two tweets *"Aren't we gonna talk about how ridiculous the new space movie with Sandra Bullock is?"* and *"I'm striving to be +ve in what I say, so I'll refrain from making a comment abt the latest Michael Bay movie"*. The first tweet contains an implicit mention of movie 'Gravity' and the second tweet contains an element of sarcasm and negative sentiment towards the movie 'Transformers: Age of Extinction'. Both the sentiment and the movie are implicit in the tweet. While it is possible to express facts, ideas, and opinions in an implicit manner, for brevity, we will focus on how knowledge aids in automated techniques for identifying implicitly mentioned entities in text. We define implicit entities as "entities mentioned in text where neither its name nor its synonym/alias/abbreviation or co-reference is explicitly mentioned in the same text".

Implicit entities are not a rare occurrence. Our studies found that 21% of the movie mentions and 40% of the book mentions are implicit in tweets, and about 35% and 40% of 'edema' and 'shortness of breath' mentions are implicit in clinical narratives. Whenever we communicate implicitly, we assume common understanding/shared-knowledge with the audience. A reader who does not know that Sandra Bullock starred in the movie 'Gravity' and that it is a space exploration movie, would not be able to decode the implicit mention of the movie 'Gravity' in the first example; a reader who does not know about Michael Bay's movie release would have no clue about the movie mentioned in the second tweet. These two examples demonstrate the indispensable value of domain knowledge in decoding implicit information in the text. State-of-the-art named entity recognition applications do not capture implicit entities [24]. Also, we have not seen big data-centric or other approaches that can identify implicit entities without the use of background knowledge (that is already available (e.g., in UMLS) or can be created (e.g., from tweets and Wikipedia)).

The task of recognizing implicit entities in the text demands comprehensive and up-to-date world knowledge. Individuals resort to a diverse set of entity characteristics to make implicit references. For example, the implicit references to the movie 'Boyhood' use phrases like *"Richard Linklater movie"*, *"Ellar Coltrane on his 12-year movie role"*, *"12-year long movie shoot"*, *"latest movie shot in my city Houston"*, and *"Mason Evan's childhood movie"*. Hence, it is important to have comprehensive knowledge about the entities to decode their implicit mentions. Another complexity is the temporal relevancy of the knowledge. The same phrase can be used to implicitly refer to different entities at different points in time. For instance, the phrase *"space movie"* could refer to the movie 'Gravity' in Fall 2013 while the same phrase in Fall 2015 would likely refer to the movie 'The Martian'. On the flip side, the most salient characteristics of the movies may change over time and so will the phrases used to refer to them. The movie 'Furious 7' was frequently referred to with the phrase *"Paul Walker's last movie"* in November 2014. This was due to the actor's death around that time. However, after the movie release in April 2015, the same entity was often mentioned through the phrase *"fastest film to reach the $1 billion"*.

We have developed knowledge-driven solutions that decode the implicit entity mentions in clinical narratives and tweets [21, 22]. Our solution models individual entities of interest by collecting knowledge about the entities from publicly available knowledge bases. These knowledge bases consists of definitions of the entities, other associated concepts, and the strength and the temporal relevance of the associated concepts. The implicit entity linking algorithms are designed to carefully use the knowledge encoded in these models to identify implicit entities in the text.

😂		🔫		💰	
Sense	**Example**	**Sense**	**Example**	**Sense**	**Example**
Laugh (noun)	I can't stop laughing 😂	Kill (verb)	He tried to kill one of my brothers last year. 🔫🔫	Costly (Adjective)	Can't buy class la 💰
Happy (noun)	Got all A's but 1😂 😊	Shot (noun)	Ooooooh shots fired! 🔫 🔫	Work hard (noun)	Up early on the grind 💰
Funny (Adjective)	Central Intelligence was damn hilarious!😂	Anger (noun)	Why this the only emotion I know to show anger? 🔫	Money (noun)	Earn money when one register /w ur link 💰

Fig. 1. Emoji usage in social media with multiple senses.

Example 2: Emoji sense disambiguation

"Emoji Sense Disambiguation" is defined as *"the machine's ability to identify the meaning of an emoji in the context in which the emoji has been used"*. This exciting new challange can benefit from carefully curated knowledge (sense inventories) and multimodal data analysis.

People are using emoji as a new language on social media to add color and whimsiness to their messages. Without rigid semantics attached to them, emoji symbols take on different meanings based on the context of a message. This has resulted in ambiguity in emoji use (see Figure 1). Similar to word sense disambiguation, machine readable knowledge bases that list emoji meanings are essential for machines to understand emoji without ambiguity. As a step towards building machines that can understand emoji, at Kno.e.sis we have developed EmojiNet [37], the first machine readable sense inventory for emoji. It links Unicode emoji representations to their English meanings extracted from the Web, enabling systems to link emoji with their context-specific meaning. EmojiNet is automatically constructed by integrating multiple emoji resources with BabelNet, which is the most comprehensive multilingual sense inventory available to date. For example, for the emoji 'face with tears of joy', EmojiNet lists 14 different senses, ranging from happy to sad. An application designed to disambiguate emoji senses can use the senses provided by EmojiNet. Emoji sense disambiguation could improve the research on sentiment and emotion analysis. For example,

consider the emoji 'face with tears of joy' which can take the meanings happy and sad based on the context in which it has been used. Current sentiment analysis application does not differentiate among the two meanings when they process the 'face with tears of joy' emoji. However, knowing the meaning of the emoji can improve sentiment prediction. Emoji similarity calculation is another task that could be benefited by knowledge bases and multimodal data analysis. Similar to computing similarity between words, we can calculate the similarity between emoji characters.

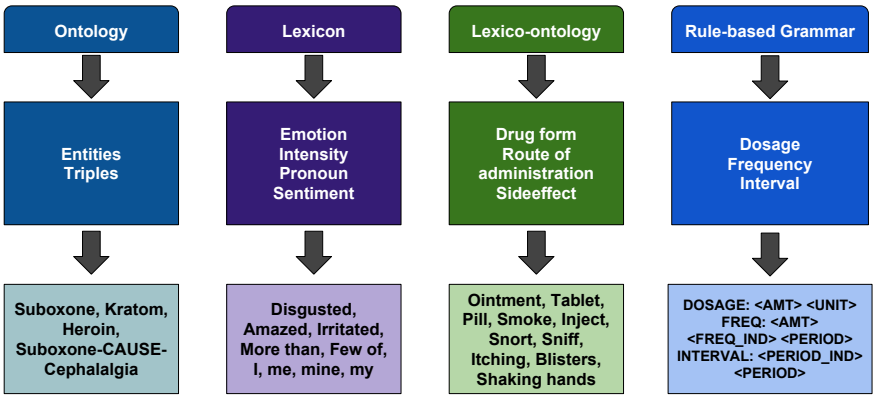

Fig. 2. Use of background knowledge to to enhance information extraction of diverse types of information. See [5] for more information.

Example 3: Understanding and analysing drug abuse-related discussions on web forums

The use of knowledge bases to improve keyword-based search has received much attention from commercial search engines lately. However, the use of knowledge bases alone cannot solve complex, domain-specific information needs. For example, to answer a complex search query such as "How are drug users engaging in the use of the semi-synthetic opioid Buprenorphine through excessive daily dosage?" may require a search engine to be aware of several facts, including Buprenorphine is a drug, users refer to Buprenorphine with synonyms such as 'bupe', 'bupey', 'suboxone', 'subbies', 'suboxone film', and the prescribed daily dosage range for Buprenorphine. The search engine should also want to have access to ontological knowledge as well as other "intelligible constructs" that are not typically modeled in ontologies, such as frequency of drug use, interval, and dosage, to answer such complex search needs. At Kno.e.sis, we have developed an information retrieval system that integrates ontology-driven query interpretation with synonym-based query expansion and domain specific rules,

to facilitate analysis of online web forums for drug abuse-related information extraction. Our system is based on a context-free grammar (CFG) that defines the interpretation of the query language constructs used to search for drug the abuse-related information needs and a domain-specific knowledge base that can be used to understand information in drug-related web forum posts. Our tool utilizes lexical, lexico-ontological, ontological, and rule-based knowledge to understand the information needs behind complex search queries (see Figures 2 and 3) [5].

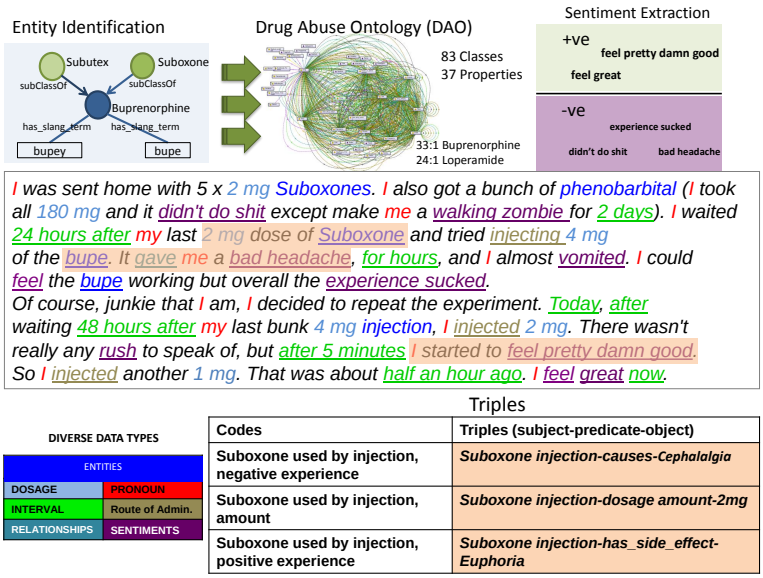

Fig. 3. Use of diverse knowledge and information extraction for deeper and more comprehensive understanding of text in health domain. See [5] for more information.

Example 4: Understanding city traffic using sensor and textual observations

With the increase in urbanization, understanding and controlling city traffic flow has become an important problem. Currently, there are over 1 billion cars on the road network, which is predicted to double by 2020, and there has been an increase of vehicular traffic by 236% from 1981 to 2001 [2]. Zero traffic fatalities and the minimization of traffic delays are some of the challenges that need to be addressed. Early understanding of traffic events is a necessity to address these challenges. The data points that help to understand such events exist in multiple modalities. Sensors deployed on road networks continuously relay important information about travel speed through certain road networks while citizen sen-

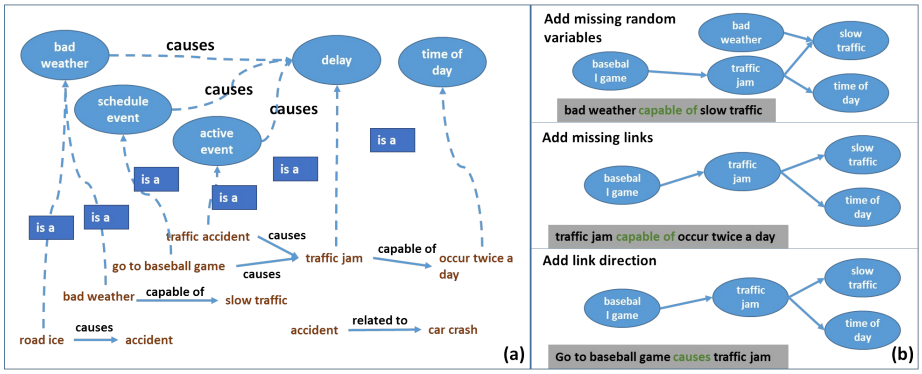

Fig. 4. (a) Domain knowledge of traffic in the form of concepts and relationships (mostly causal) from ConceptNet (b) Probabilistic graphical model (PGM) that explains the conditional dependencies between variables in traffic domain (only a part is shown in the picture) is enriched by adding the missing random variables, links and link directions extracted from ConceptNet. Figure 5 shows how this enriched PGM model is used to correlate contextually related data of different modalities. See [3] for more information.

sors (i.e. humans) share real-time information about traffic/road conditions on public social media streams such as Twitter. As humans, we know how to integrate information from these data sources and understand traffic events (i.e. the slow-moving traffic shown by sensor observations could be due to an accident reported in tweets at location x). However, current research on understanding city traffic dynamics focuses only on either sensory data or social media data. We can exploit the complementary and corroborative nature of these data sources to understand the traffic events.

The first step towards such effort is to materialize the domain knowledge possessed by humans about traffic events to a machine readable format (i.e. to develop a knowledge base on traffic events and their causes). A statistical approach would help to establish the associations between variables of the domain (e.g. there is an association between 'bad weather' and a 'traffic jam'). However, such approaches fall short in finding all the variables that exist in the domain, finding all associations between variables, and finding the causal directionality between variables. We developed techniques to leverage domain knowledge to enrich the statistical models that address above shortcomings. Primarily we used the sensor data collected by 511.org to develop an initial probabilistic graphical model that explains the conditional dependencies between variables in traffic domain. Then we leverage the domain knowledge encoded in ConceptNet to add more nodes to the model that represent missing variables, add more edges to the model that represent missing associations, and add directionality to the edges between nodes to represent the conditional dependencies. Figure 4(a) shows a snippet of the ConceptNet and Figure 4(b) demonstrates the enrichment step of the developed model using the domain knowledge in ConceptNet [3].

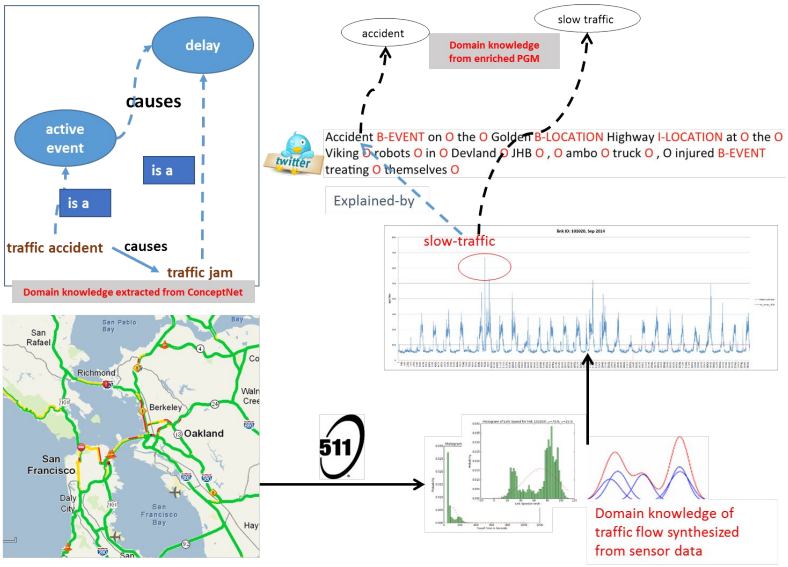

Fig. 5. Use of domain knowledge to correlate contextually (temporal, spatial) related data of different modalities (sensor and tweet). In this case, slow moving traffic at a geo-location and time is correlated with a tweet indicating an accident event at a location entity described in the text. See [3] for more information.

The next step is to understand the sensor observations. The idea is to model 'normal' traffic patterns using sensor observations and then detect any anomalies. We used a Restricted Switching Linear Dynamical System (RSLDS) to model normal speed and travel time dynamics and thereby characterize anomalous dynamics. Using speed and travel time data from each link, we generated time series data for each hour of the day and for each day of the week. Then for each hour of each day of week, we averaged the travel speeds and travel times at each road link (see Figure 4). However, the average speed would not be a real speed that we observed in the data. Thus, to select a speed that exists in the actual data we chose the speed that is closest to the average speed using a point-wise Euclidean distance metric. We used the above speed data to learn parameters for a RSLDS (using log likelihood function), and treated the RSLDS model as a model for the normal traffic dynamics of the San Francisco Bay Area [2]. If the log likelihood value for a particular day of the week and hour of the day is less than the minimum log likelihood value for that time period, we tagged the traffic dynamics as anomalous.

The anomalous observations are further analyzed with events extracted from Twitter data and subsequently declared as being explained when we can find traffic causing events in Twitter data (according to the domain model) with a time overlap of the observed anomalous traffic behaviour. Figure 5 demonstrates this process. This example again demonstrates vital role domain knowledge plays

to improve the ability of a AI technique such as RSLDS (that help to process large amounts of data to develop a model and identification of anomaly) to improve interpretation of data and in integrated exploitation of complementary data of different modalities. Further exploration of different approaches to represent and exploit semantics appears in [32].

3 Conclusions and looking forward

We discussed the importance of domain/world knowledge in understanding complex data in the real world, particularly, when large amounts of training data are not readily available or is expensive to generate. We demonstrated two use cases (and referred to three additional applications) where knowledge plays an indispensable role in understanding complex language constructs and multimodal data. We are also seeing early efforts in making knowledge bases dynamic and evolve to account for the changes in the real world [33].

Knowledge seems to play a central role in human learning and intelligence, such as in learning from small amount of data, or in cognition – especially perception. Our ability to create or deploy just the right knowledge in our computing processes will improve machine intelligence, perhaps in a similar way as knowledge plays a central role in human intelligence. As a corollary of this, two of the specific advances we will see are a deeper and nuanced understanding of content (including but not limited to text) and our ability to process and learn from multimodal data at a semantic level (given that concepts manifest very differently at the data level in different media or modalities). The human brain is extremely adept at processing multimodal data – our senses are capable of receiving 11 million bits per second, and our brain is able to distill that into a few tens of bits of abstractions (for further explorations, see [29]). Knowledge plays a central role in this abstraction process known as the perception cycle.

Machine intelligence has been the holy grail of a lot of AI research lately. The statistical pattern matching approach and learning from big data, typically of single modality, has seen tremendous success. For those of us who have pursued brain-inspired computing approaches, we think the time has come for rapid progress using a model-building approach. The ability to build broad (both in terms of coverage as well as variety- not just entities and relationships, but also emotions, intentions and subjectivity features; linguistic, cultural and other aspects of human interest and functions) and static knowledge to domain-specific, purpose-specific, personalized, and/or dynamic knowledge combined with richer representation – especially probabilistic graph models – will see very rapid progress. These will complement neural network approaches. We may also see knowledge playing a significant role in enhancing deep learning. Rather than the dominance of data-centric approaches, we will see an interleaving and interplay of the data and knowledge tracks.

Acknowledgments

We acknowledge partial support from the National Science Foundation (NSF) award: CNS-1513721: "Context-Aware Harassment Detection on Social Media", National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02: "Trending: Social Media Analysis to Monitor Cannabis and Synthetic Cannabinoid Use", National Institutes of Health (NIH) award: MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression", and Grant No. 2014-PS-PSN-00006 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the U.S. Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice, NSF, NIH or NIDA.

References

1. Anantharam, P., Banerjee, T., Sheth, A., Thirunarayan, K., Marupudi, S., Sridharan, V., Forbis, S.G.: Knowledge-driven personalized contextual mhealth service for asthma management in children. In: 2015 IEEE International Conference on Mobile Services. IEEE (2015)
2. Anantharam, P., Thirunarayan, K., Marupudi, S., Sheth, A., Banerjee, T.: Understanding city traffic dynamics utilizing sensor and textual observations. In: Proceedings of The 13th AAAI Conference on Artificial Intelligence (AAAI-16), February 12–17, Phoenix, Arizona, USA (2016)
3. Anantharam, P., Thirunarayan, K., Sheth, A.P.: Traffic analytics using probabilistic graphical models enhanced with knowledge bases. Analytics for Cyber Physical Systems workshop at the SIAM Conference on Data Mining (SDM) (2013)
4. Balasuriya, L., Wijeratne, S., Doran, D., Sheth, A.: Finding street gang members on twitter. In: The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016). vol. 8. San Francisco, CA, USA (2016)
5. Cameron, D., Sheth, A., Jaykumar, N., Thirunarayan, K., Anand, G., Smith, G.A.: A hybrid approach to finding relevant social media content for complex domain specific information needs. Web Semantics: Science, Services and Agents on the World Wide Web 29 (2014)
6. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R., et al.: What are ontologies, and why do we need them? IEEE Intelligent systems 14(1) (1999)
7. Cruse, D.A.: Lexical semantics. Cambridge University Press (1986)
8. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM 55 (2012)
9. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al.: Building watson: An overview of the deepqa project. AI magazine 31(3) (2010)
10. Fillmore, C.J.: Frame semantics and the nature of language. Annals of the New York Academy of Sciences 280(1) (1976)

11. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems 24 (2009)
12. Jadhav, A.: Knowledge Driven Search Intent Mining. Ph.D. thesis, Wright State University (2016)
13. Kimmig, A., Bach, S., Broecheler, M., Huang, B., Getoor, L.: A short introduction to probabilistic soft logic. In: Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications (2012)
14. Lalithsena, S., Hitzler, P., Sheth, A., Jain, P.: Automatic domain identification for linked open data. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies. vol. 1 (2013)
15. Meng, L., Huang, R., Gu, J.: A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology 6(1) (2013)
16. Mihalcea, R.: Knowledge-based methods for wsd. Word Sense Disambiguation: Algorithms and Applications (2006)
17. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. International journal of lexicography 3(4) (1990)
18. Nguyen, V., Bodenreider, O., Sheth, A.: Don't like rdf reification?: making statements about statements using singleton property. In: Proc. of the 23rd inti. conf. on WWW (2014)
19. Ovchinnikova, E.: Integration of world knowledge for natural language understanding, vol. 3. Springer Science & Business Media (2012)
20. Perera, S., Henson, C., Thirunarayan, K., Sheth, A., Nair, S.: Semantics driven approach for knowledge acquisition from emrs. IEEE journal of BHI 18(2) (2014)
21. Perera, S., Mendes, P., Sheth, A., Thirunarayan, K., Alex, A., Heid, C., Mott, G.: Implicit entity recognition in clinical documents. In: Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (* SEM) (2015)
22. Perera, S., Mendes, P.N., Alex, A., Sheth, A., Thirunarayan, K.: Implicit entity linking in tweets. In: International Semantic Web Conference. Springer (2016)
23. Ramakrishnan, C., Mendes, P.N., da Gama, R.A., Ferreira, G.C., Sheth, A.: Joint extraction of compound entities and relationships from biomedical literature. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (2008)
24. Rizzo, G., Basave, A.E.C., Pereira, B., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.: Making sense of microposts (# microposts2015) named entity recognition and linking (neel) challenge. In: # MSM. pp. 44–53 (2015)
25. Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R., Scheffczyk, J.: Framenet ii: Extended theory and practice (2006)
26. Shekarpour, S., Ngonga Ngomo, A.C., Auer, S.: Question answering on interlinked data. In: Proceedings of the 22nd international conference on World Wide Web (2013)
27. Sheth, A.: 15 years of semantic search and ontology-enabled semantic applications. Blog – http://j.mp/15yrsSS (2014)
28. Sheth, A., Anantharam, P., Henson, C.: Physical-cyber-social computing: An early 21st century approach. IEEE Intelligent Systems 28(1) (2013)
29. Sheth, A., Anantharam, P., Henson, C.: Semantic, cognitive, and perceptual computing: Paradigms that shape human experience. Computer 49(3) (2016)
30. Sheth, A., Avant, D., Bertram, C.: System and method for creating a semantic web and its applications in browsing, searching, profiling, personalization and advertising (Oct 30 2001), uS Patent 6,311,194

31. Sheth, A., Kapanipathi, P.: Semantic filtering for social data. IEEE Internet Computing 20(4) (2016)
32. Sheth, A., Ramakrishnan, C., Thomas, C.: Semantics for the semantic web: The implicit, the formal and the powerful. International Journal on Semantic Web and Information Systems (IJSWIS) 1(1), 1–18 (2005)
33. Sheth, Amit Thomas, C., Mehra, P.: Continuous semantics to analyze real-time data. Wiki – http://bit.ly/2cVGbov (2010)
34. Shoham, Y.: Why knowledge representation matters. Communications of the ACM 59(1) (2015)
35. Thirunarayan, K., Sheth, A.: Semantics empowered big data processing with applications. AI Magazine 36 (2015)
36. Wang, W.: Automatic Emotion Identification from Text. Ph.D. thesis, Wright State University (2015)
37. Wijeratne, S., Balasuriya, L., Sheth, A., Doran, D.: Emojinet: Building a machine readable sense inventory for emoji. In: 8th International Conference on Social Informatics (SocInfo 2016). Bellevue, WA, USA (2016)
38. Winograd, T.: Understanding natural language. Cognitive psychology 3(1) (1972)