SYSTEMS ENGINEERING RESEARCH CENTER

CENTER FOR COMPLEX SYSTEMS & ENTERPRISES

STEVENS INSTITUTE OF TECHNOLOGY 1870

Research and Application Workshop

AI4SE & SE4AI

NLP and Knowledge Engineering to extract models from text

Carlo Lipizzi – clipizzi@stevens.edu

September 2022

The context

- Models* provide a way to solve real-world problems safely and efficiently. They are an important method of analysis which is easily verified, communicated, and understood. We use them when conducting experiments on a real system is impossible or impractical, often because of cost or time [AnyLogic]
- Models are never as good as the reality. They are as good as their representation of the system they model
- They have embedded the knowledge we have of the system we want to represent
- The more accurate and comprehensive is the knowledge and its representation, the more accurate => useful the model is
- This has merit in traditional modeling as well as AI/ML models and applications as digital twins

*model is a physical, mathematical or logical representation of a system, entity, phenomenon, or process [SYS 611]

The study of Knowledge - Epistemology

- Epistemological questions explore the nature of knowledge

  Ask how someone has come to know something, inquire into the scope and limits of knowledge or try to discover the degree of certainty attached to particular knowledge

- E.g.: the stick that appears to bend in the water

  Use the knowledge of science to rationalize that the stick is not bent but it is the refraction of light in the water that makes it look that way

  But the epistemologist might ask how do you really know that the stick does not actually bend in the water

Why epistemology is relevant for AI?

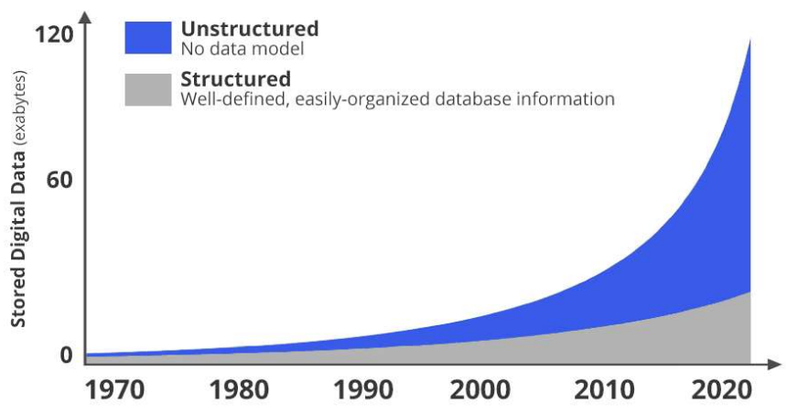
- We cannot represent what we don’t know
- If we don’t/cannot fully “know”, we should have an approximation of the knowledge, supported by theoretical and empirical evidences, possibly knowing its limitations
- Epistemology is studying knowledge, advocating models to represent it
- We should use epistemology models as input for the mathematical models to be used to write the code for our AI systems
- Without a solid framework for representing knowledge, we may face the risk of a new “AI winter”

Philosophy and AI/ML

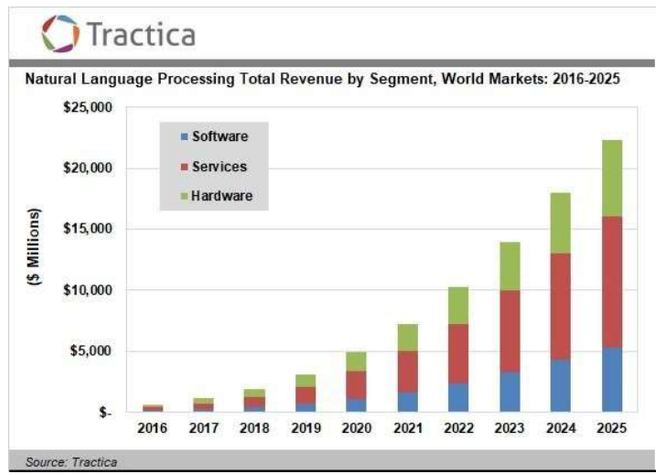
Philosophy	AI/ML
▪ **Rationalists** - Believe that knowledge comes from exercising the human ability to reason. Reason not only enables people to know things that the senses do not reveal but it is also the primary source of knowledge. Plato and Descartes were rationalists	▪ **Symbolic Reasoning – traditional AI** - Using preset symbolic structure to get knowledge about a given problem. Taxonomies, ontologies, rules (IF/THEN) are examples
▪ **Empiricists** - Believe that knowledge comes from experience. This is evidence provided by the senses	▪ **Data Driven – Machine Learning** - Applying algorithms to large collection of data “describing” the reality to be represented. This is in line with advanced statistical models, centered on pattern recognition

Natural Language as source of Data

- 85-90 percent of all corporate data is in some kind of unstructured form, such as text and multimedia [Gartner, 2019]
- Tapping into these information sources is a need to stay competitive



Source: m-files.com



- Text conveys a great portion of the knowledge people have about a given domain

Implementing NLP/NLU

- Language is changing constantly, and NLP is following the changes, going from processing based on predefined structures (taxonomies/ontologies, syntax) to structures deducted from the text itself

Limitations of the traditional-deductive-"symbolic" approach

- Today, language is more fragmented, has less structure, has more jargons
- Different points of view may provide different interpretations

Machine Learning/inductive approach

- Extracting a numerical structure from text
- Different structures for different points of view
- Different structures automatically extracted over time

Testing the two approaches

- In order to compare the 2 approaches, we defined 2 tasks:
  - Named Entity Recognition (NER)
  - Semantic Role Labeling (SRL). We are using SPO semantic triples (subject, predicate, object)

  We selected those 2 tasks because they are essential building block of most of the models
- To make the comparison more accurate, we are working on 2 types of documents:
  - General purpose
  - Domain specific

  For each one of the 2 types, we are analyzing a longer and smaller documents

Documents	Total #sentences	Total #words
Long_generic (HP)	6480	77290
Short_generic (NYtimes)	54	1121
Long_domain-specific (Neurology)	13719	235093
Short_domain-specific (Brain Inflammation)	712	8464

The tools

Symbolic approach

- **spaCy**: To create customized training dataset for data-driven approach
- **coreNLP**: To extract NEs and SPO triples
- **NLTK**: We use nltk.wordnet to capture taxonomy structure (hypernym-hyponym) of extracted named entity and SPO semantic triples

Data-driven approach

- **XLNet**: this is a Transformer-based generalized NLP tool (*Generalized Autoregressive Pretraining for Language Understanding*)

Preliminary results

- Evaluating semantic accuracy in text needs to have **humans** involved. Automatic evaluation would be affected by the bias in the annotated text used for the evaluation
- So far, we run NER with both the approaches, SRL with the symbolic one only
- The **data-driven** approach has very good results on large generic datasets, poor results on domain-specific and smaller in particular. This is because those models use as semantic base for pre-training large generic texts. It could be possible to use customized semantic bases, but they should be sizable. Pre-training on reasonable large datasets could take weeks of computing time
- The **symbolic** approach is as good as the “symbols” we use. We used general ones, with good results on the large generic documents (but worst than the data-driven), good on small generic documents (better than the data-driven). Inconclusive the results on the domain-specific

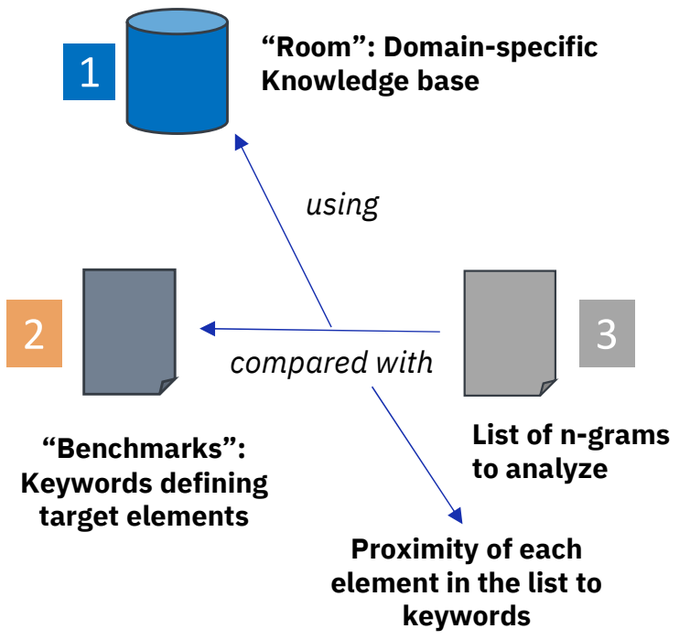
Results interpretation

- The **data-driven** approach is using a “mechanical” approach to semantic that does not reflect the way we develop and use our knowledge. The underlying theoretical method is correlation/pattern recognition, but we reason in more complex ways. The complexity of the algorithms inside those models makes understanding what is going inside practically impossible: the vast number of layers of the neural networks inside would require a memory of the status of each layer and this is not there. In theory, this approach could be totally unsupervised (with pre-definable bias), but the cost of pre-training makes this option non applicable
- The **symbolic** approach is as good as the symbols (taxonomies, rules, meta languages) that are used. Symbols are domain-specific and may change in time. This is a fully supervised approach
- What is missing is a model representing the knowledge, able to use algorithms and approaches as its components

Moving forward

- While we will complete the comparison of the 2 traditional approaches (data-driven and symbolic), we will introduce a method based on a knowledge representation model we developed (the “room theory”) and using graph theory
- The ”room theory” is a framework to address the relativity of the point of view by providing a computational representation of the context
- The non computational theory was first released as “schema theory” by Sir Frederic Bartlett (1886–1969) and revised for AI applications as “framework theory” by Marvin Minsky (mid ‘70)
- For instance, when we enter a physical room, we instantly know if it is a bedroom, a bathroom, or a living room
- Rooms/schemata/frameworks are mental frameworks we use to organize remembered information and represent an individuals/domain-specific view of the reality

How the “room theory” works

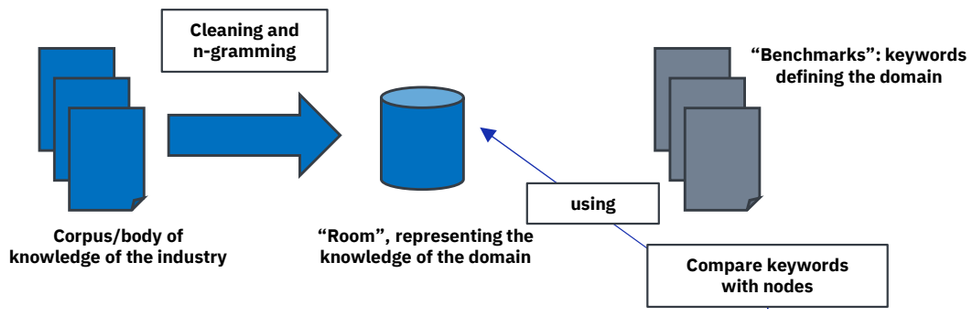


- “Room theory” enables the use of context-subjectivity in the analysis of the incoming documents
- Context-subjectivity can be the point of of view of a subject matter expert
- The context-subjectivity in the analysis is represented by a domain specific numerical knowledge base, created from a large domain specific & representative corpus that is then transformed into a numerical dataset (“embeddings table”)

- The key components are:

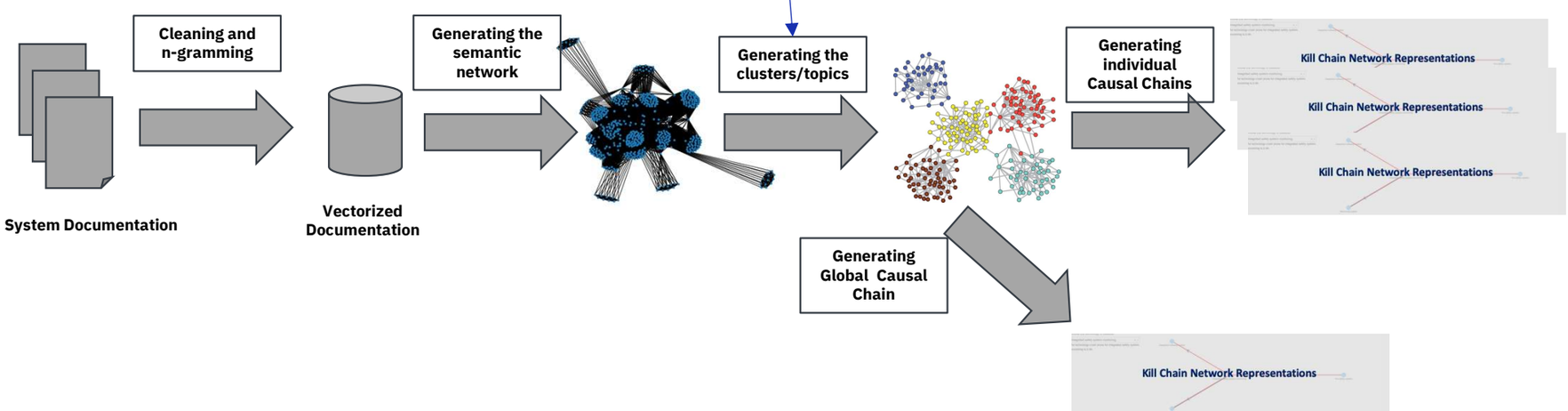
1. A point of view for the comparison (the “room”). This is represented by the embeddings table extracted from a large/representative corpus from the specific domain
2. A list of “extended” keywords (using synonyms and misspellings) to be used for the analysis (the ”benchmark”)

Our Approach – putting things together



- The flow on top provides adjustments based on domain-specific knowledge

- The flow at the bottom is the actual workflow to get NERs (from the clusters) and SRLs (from the network)



SYSTEMS ENGINEERING RESEARCH CENTER

Thank you!

CENTER FOR COMPLEX SYSTEMS & ENTERPRISES

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Dr. Carlo Lipizzi
clipizzi@stevens.edu

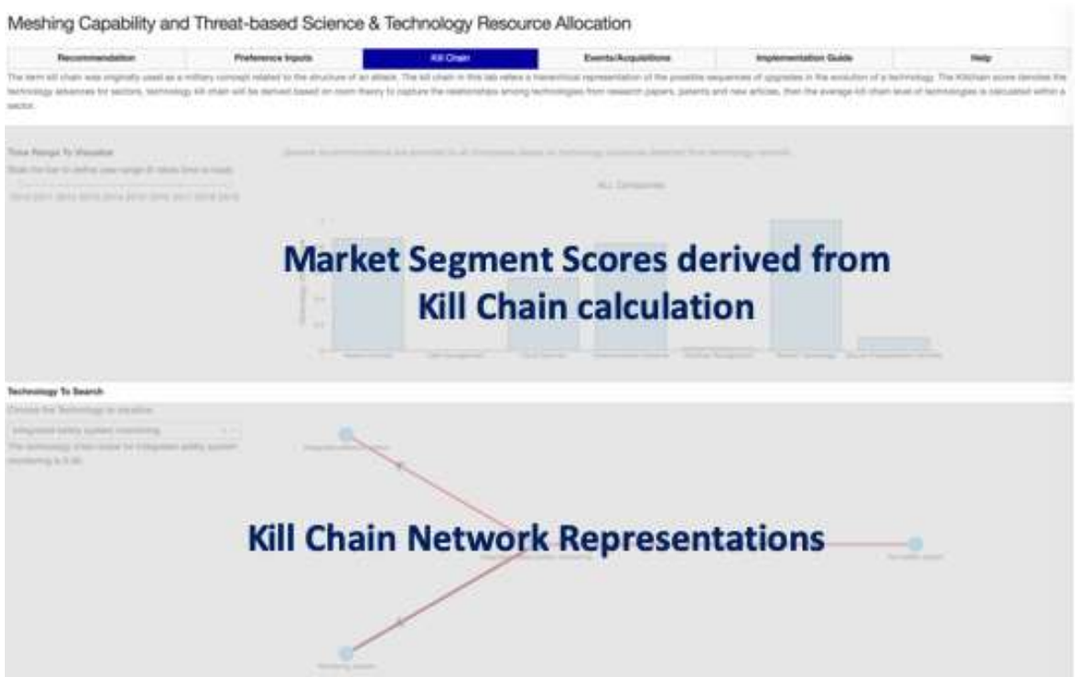
Shiyu Yuan – ***PhD Candidate***
syuan14@stevens.edu

Our Approach – putting things together

- We prune the list of ngrams using the room theory
- We create ego networks for the “subjects”. The degrees of separation is function of the size of the cluster
- The ego networks represent the semantic dependency between the nodes within the topics
- The approach can be extended to inter-clusters relations to recreate the complete formal representation

How we use it so far

- We used it to determine the causal chain in the domain of technologies
- Each technology has “components”, that are other technologies required for the first one. For example, cell. phones <- batteries, display, antennas, …



- The model has been partially implemented in WRT-1010 “Meshing Capability and Threat-based Science & Technology Resource Allocation”