Semantic Information Extraction and Generation of Dynamic Knowledge Graphs

Damir Cavar

February 2019

University of Illinois at Urbana-Champaign

Agenda

- Goals
- Knowledge Graphs
- Information Extraction
- NLP now and then
- Issues
- HooSIER Knowledge Graph Extractor
- Demo

Goals

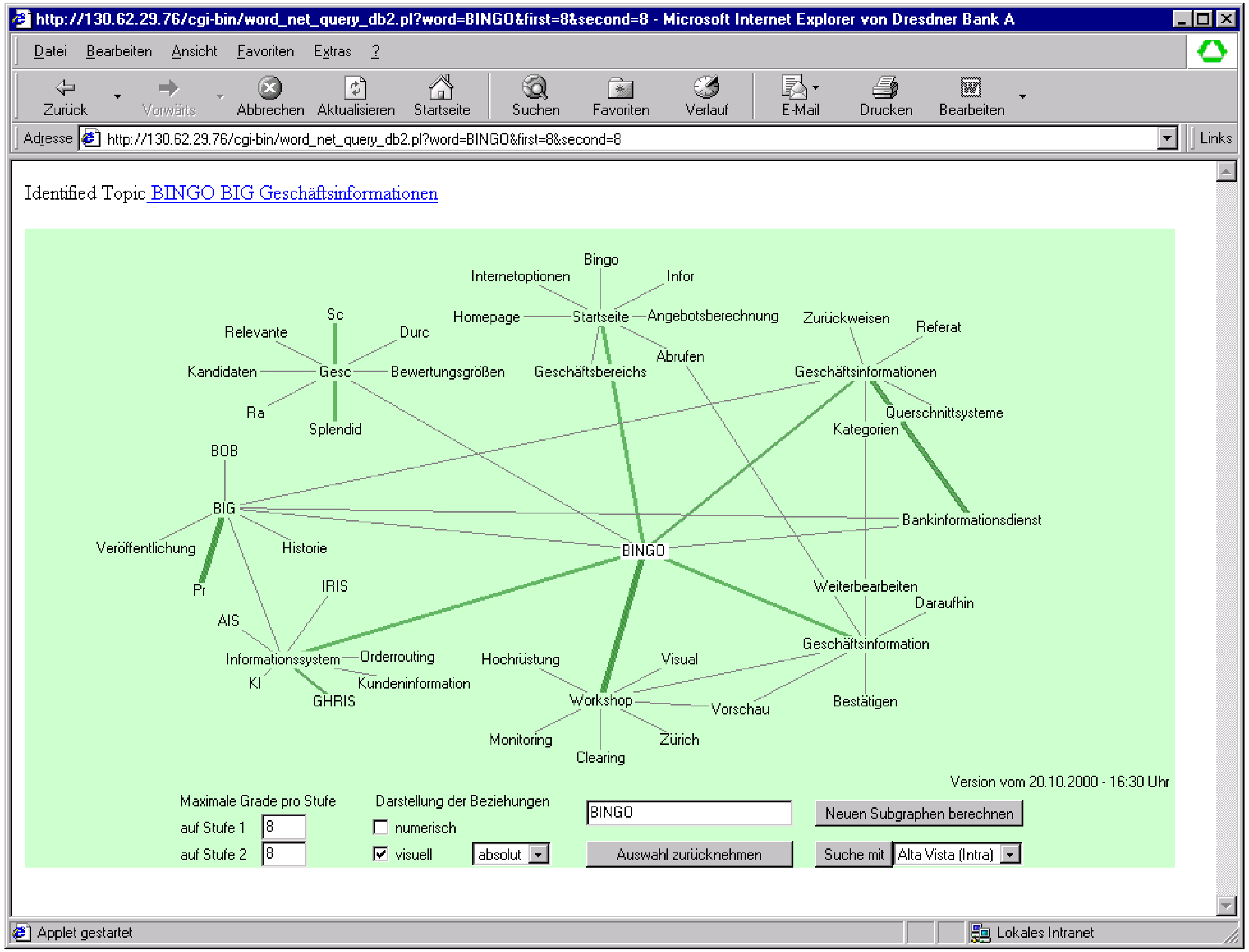
- Information Extraction:
    - Entities and Relations from text
        - Open domain and domain specific
    - Description of concepts, relations, detailed semantic properties using
        - Description Logic approach
        - Knowledge Graph approach
        - Linking and Typing of entities and relations
- Natural Language Processing:
    - Semantic and Pragmatic processing
        - Implicatures and Presuppositions
        - Reasoning and Common Sense
    - Linguistic Processing
- Scalable and High-Performance Big-Data NLP for Text 2 Data

Knowledge Graphs

- Assumption:
  - First mention of term in a Google Blog
    - Amid Singhal (2012), Introducing the Knowledge Graph: things, not strings https://www.blog.google/products/search/introducing-knowledge-graph-things-not/
- Reality:
  - Use of Graphical Knowledge Representation is older
    - Description Logic
    - RDF, OIL and DAML to OWL
    - Applications

Knowledge Graphs back in 2000

- RDB-based SemNet
  - Prior to OWL
  - OIL, DAML were around
  - No GraphDB
  - No NLP technologies (Stanford CoreNLP, OpenNLP, spaCy, Polyglot, GATE, etc.)

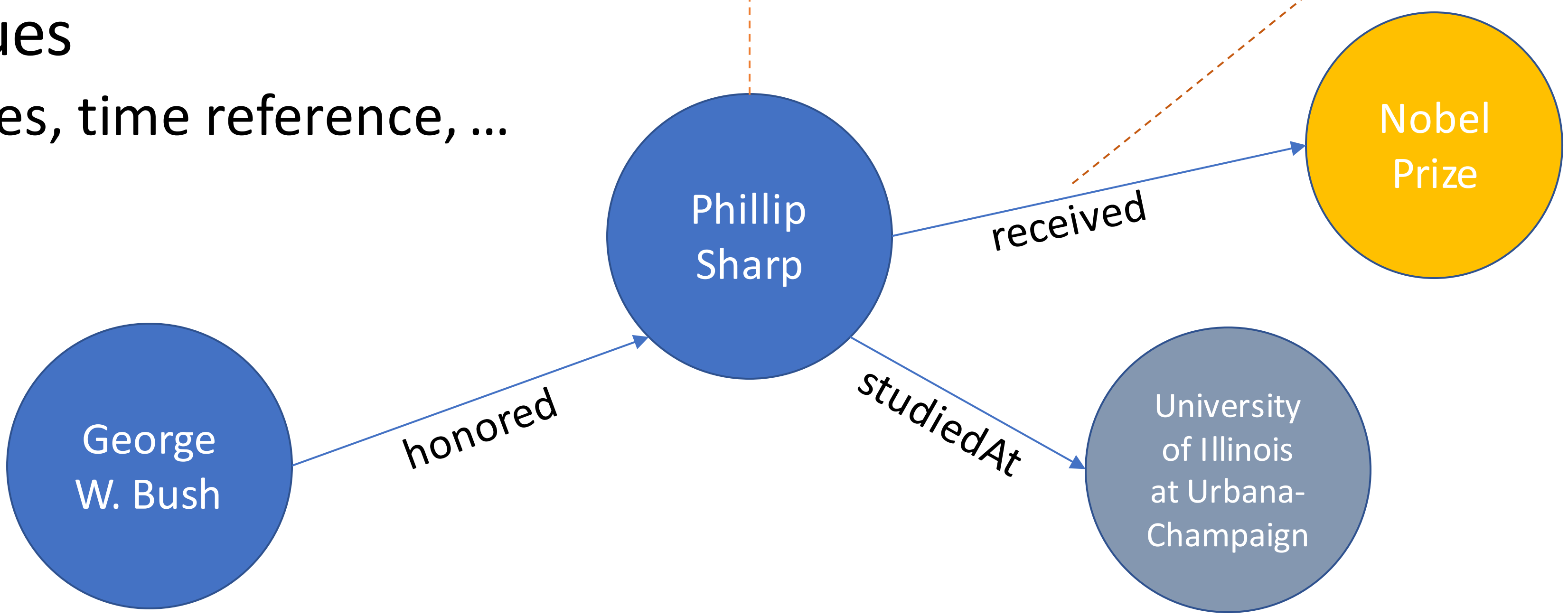


Knowledge Graphs

- Concepts and Relations
  - Mostly unconstrained
  - Domain specific or free
- Attributes and Values
  - encoding properties, time reference, …

Attribute	Value
Birth place	Falmouth, KY
DOB	06/06/1944
gender	male
…	

Attribute	Value
year	1993
…	

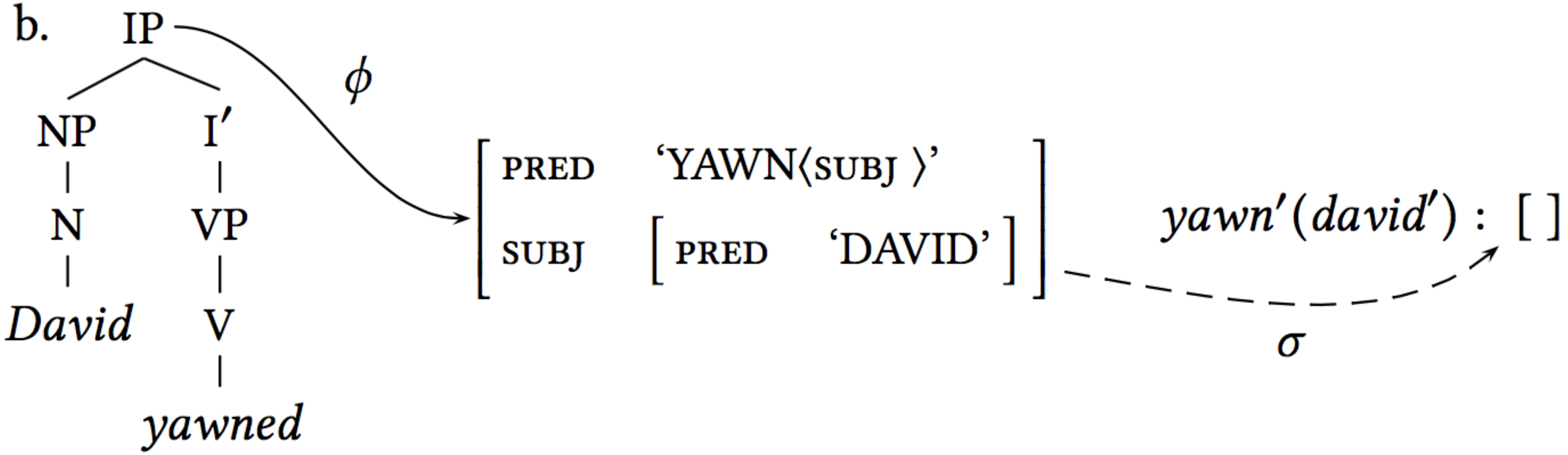


Formal Semantics

- Meaning and Compositionality as Formal Mapping from Syntax to Semantic Representation

a. David yawned.

b.



Knowledge Graphs

- No computation or interpretation of logic equations
- Direct mapping of knowledge from text

- Description of Knowledge
  - Directed Graph: encoding concept, events, domain specific knowledge…
  - Attribute-Value encoded features like size and shape, but also event time references (start, end, duration), etc.

- Reasoning
- Prediction
- Machine Learning of concepts and concept properties

State-of-the-Art

- Information Extraction
  - Open IE
  - Language Agnostic IE
    - Entity detection
    - Entity-Relation extraction
- Knowledge Graphs and Knowledge Representations
  - Ontology learning
  - Entity and Relation Linking

OpenIE

- Unstructured natural language expressions to structured representations (Banko et al., 2007)
  - Structured representation:
    - Relational tuples of semantic relations: argument – predicate – argument
    - Relations are not a priori specified (not domain specific)
    - Extraction of all entities and relations
    - Domain agnostic entity and relation discovery
- Example:
  - Tim Cook, the CEO of Apple and a board member of Alphabet Inc., announced that he will no longer serve in any function for Apple Inc.

OpenIE

- Underlying goal:
  - Tim Cook, the CEO of Apple and board member of Alphabet Inc. (…)
    - Tim Cook – isA – CEO of Apple
    - Tim Cook – isA – board member of Alphabet Inc.
    - Not in the last relation ignored completely!
- Reality:
  - Tim Cook – CEO of – Apple
    - No relation to Alphabet Inc.
  - He – serve in – function for Apple Inc.
    - No anaphora resolution
    - No processing of Negation

OpenIE Issues

- Underlying NLP ranks between “acceptable” and “of limited use at best.”
- Entity recognition is broad
  - Coreference analysis not reliable
- Lack of Linking
  - Entities identified via Linking to concepts in Knowledge Graphs (e.g. YAGO, DBpedia)

NLP Technologies

- Back in 2000
  - Regular expressions and pattern matching
  - Template-based text generation
  - Finite State Dialog modeling
  - Knowledge Graphs (SemNets) on RDBs
  - Text2Speech

  - Part-of-speech tagging
  - Parsing
  - Machine Translation

- Rule-based systems, probabilistic models, knowledge-based NLP

NLP Technologies

- 2019: Focus on limited model types and technologies:
  - Data driven and usage based modeling, ignoring knowledge, rules, universals

  - Dependency Parse Trees from treebanks
  - Treebank-derived Constituent Tree Parsers
  - Label/Tag-based Semantic Role Labeling
  - …
  - Pipeline-architecture as such:
    - Isolated modules with very limited NLP-focus chained in an input-output pipeline
      - CoreNLP, spaCy, OpenNLP, LingPipe, GATE, NLTK, UIMA, …

  - No parallel architectures!

NLP Technologies

- State of the Art: (Sebastian Ruder’s overview)
  - Part-of-Speech Tagging:
    - Use: word-level part of speech annotation with a limited set of tags that encode some morphosyntactic features
    - **F1 score**: **95% - 97%** based on WSJ portion of Penn Treebank, more than 100 treebanks for UD
    - Best performing: Deep Learning Approaches (alternatives not evaluated)

NLP Technologies

- State of the Art: (Sebastian Ruder's overview)
  - Constituent Tree Parsing:
    - Use: phrasal structure; relations, hierarchies and ambiguities between phrases; semantic scope relation; …
    - **F1 score**: **92% - 95%** based on Penn Treebank
    - Best performing: Deep Learning Approaches (alternatives not evaluated)
  - Dependency Parsing:
    - Use: dependency relations between elements in the sentence; simplified annotation of functional relations: Subject, Object, Modifier, …
    - **F1 score** on labels and relations: **91% - 94%** based on Stanford Dependency conversion of the Penn Treebank
    - Best performing: Deep Learning Approaches (alternatives not evaluated)

NLP Technologies

- State of the Art: (Sebastian Ruder's overview)
  - Named Entity Recognition:
    - Use: entity labeling – person, institution, location, time, currency, …
    - **F1 score**: **90% - 92%** based on Reuters RCV1 corpus with **four** NE-types (PER, LOC, ORG, MISC) using BIO notation
    - Best performing: Deep Learning Approaches (alternatives not evaluated)
  - Semantic Role Labeling:
    - Use: Label predicate argument structure (*Who gave what to who*): Predicate, Subject, Object, entity and relation extraction
    - F1 score: **81% - 84%** based on OntoNotes benchmark of the Penn Treebank
    - Best performing: Deep Learning Approaches (alternatives not evaluated)

NLP Technologies

- F1 score margins and error rates:
  - Basic token-level classification: error of approx. **4%**
  - Word-level annotation, syntactic parsing: **10%**
  - Semantic-level annotation: **30%**

- What has changed since 2000?
  - Cross-linguistic Coverage
  - Speed

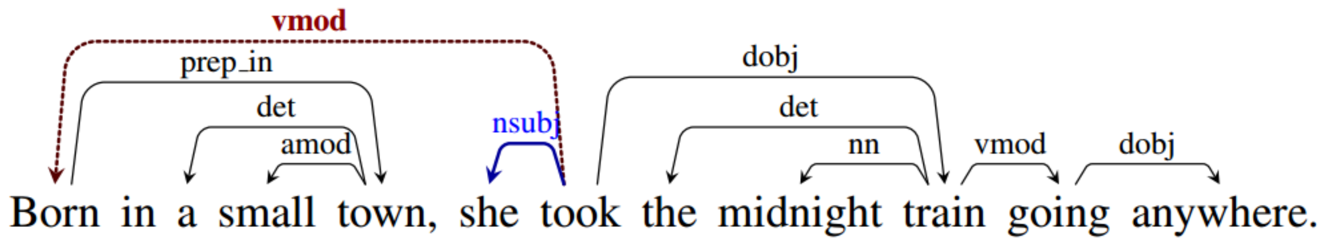
- Situation check:
  - Mono-culture of training/test-datasets for data driven ML/DL-methods
  - Limitation to weak linguistic models (e.g. *Constituent Trees*, *NE-classes*, *Semantic roles*), annotation standards (e.g. *Dependencies*)

NLP Technologies

- Situation check:
  - Limited use of NLP-pipelines: PoS-tagging, Lemmatization
    - CoreNLP: Constituent Parser; Dependency Parser; Coreference Analysis; …
    - spaCy: Dependency Parser
    - NLTK: WordNet
  - Lack of APIs that interface to linguistic output data structures
    - NLP developers lack understanding of the linguistic annotations generated by pipelines or tools

NLP Example

- Stanford Open IE (paper and website)



  - Lack of intuition of dependency relations
    - Modification of ROOT (took) by “born in a small town” is counterintuitive
- Lack of:
  - Clause level hierarchical relation analysis (subordinate clauses and scope)
  - Tempus, Mood, … annotation
  - Pragmatic and semantic properties (and relevant linguistic features)

Issues

- Transparency
  - Lack of understanding of linguistic annotations
  - No abstraction layer and API
  - Blackbox models without introspection
    - Deep Learning
- Data-driven Systems
  - Knowledge driven engineering impossible
    - Lacking grammar engineering interface
  - Large data sets necessary
  - Monoculture of data sets
- Error rate in a pipeline

Issues

- NLP Technologies and Language Resources
  - More than 7,100 estimated languages
  - 300 estimated to be written
  - 1% is well resourced (data and technology wise)
- Language Resources
  - Mono-culture
    - Limited data set or corpora as “standard”
    - Evolutionary model of technologies that are tuned to excel on the “standard”
  - Half-life of resources
    - Corpora use value
  - Annotation
    - Errors
    - Theoretically motivated

NLP Example

- Scope between clauses:
  - Reuters reported [that [Google bought Apple]]
  - Reuters did not report [that [Google bought Apple]]
  - Reuters did not deny [that [Google bought Apple]]
- Tense:
  - Tim Cook bought Google.
  - Tim Cook will buy Google one day.

NLP Technologies

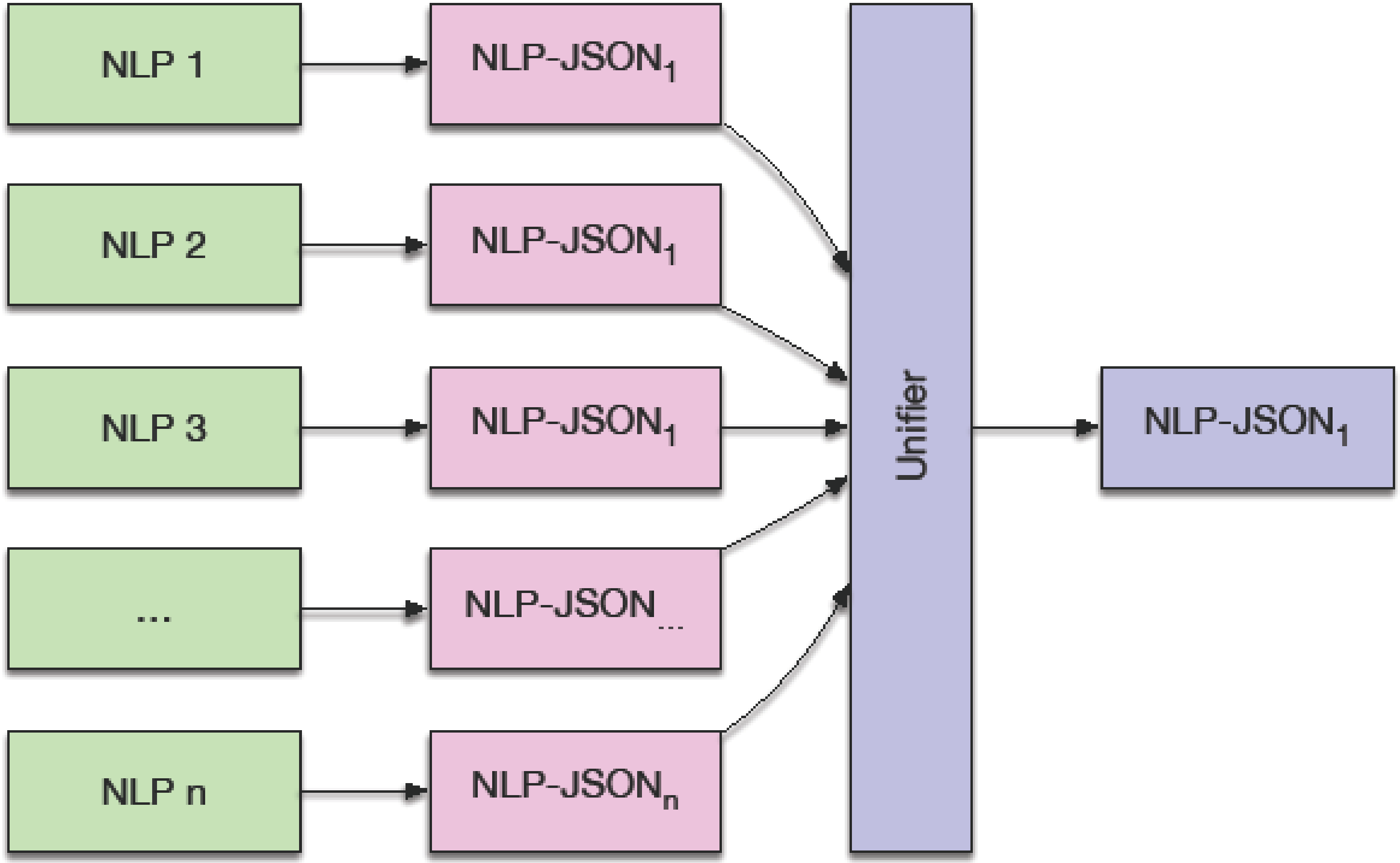
- Applied to real text:
  - Sentence length over 10 to 15 tokens breaks common probabilistic or NN parsers (Dependency parsers, in particular)
- Problematic domains, for example:
  - SEC, Financial, or Business Reports
  - Case-law and legal documents
  - Medical text (patient reports, documentations)
- Current free and open NLP-pipelines are of limited use.
- Are they of any use for serious NLP-based technologies?

State of the Art

- Δ between 2000 – 2018
  - ASR improvements
  - Knowledge Graphs, Ontologies
  - Integration
    - Data sources
    - Interfaces, multi-modal interaction
    - Device architecture
- Is there any significant progress in ___ ?
  - Dialog management
  - NLP at the utterance and discourse level
  - Semantics and Pragmatics

NLP Ensemble

- HooSIER

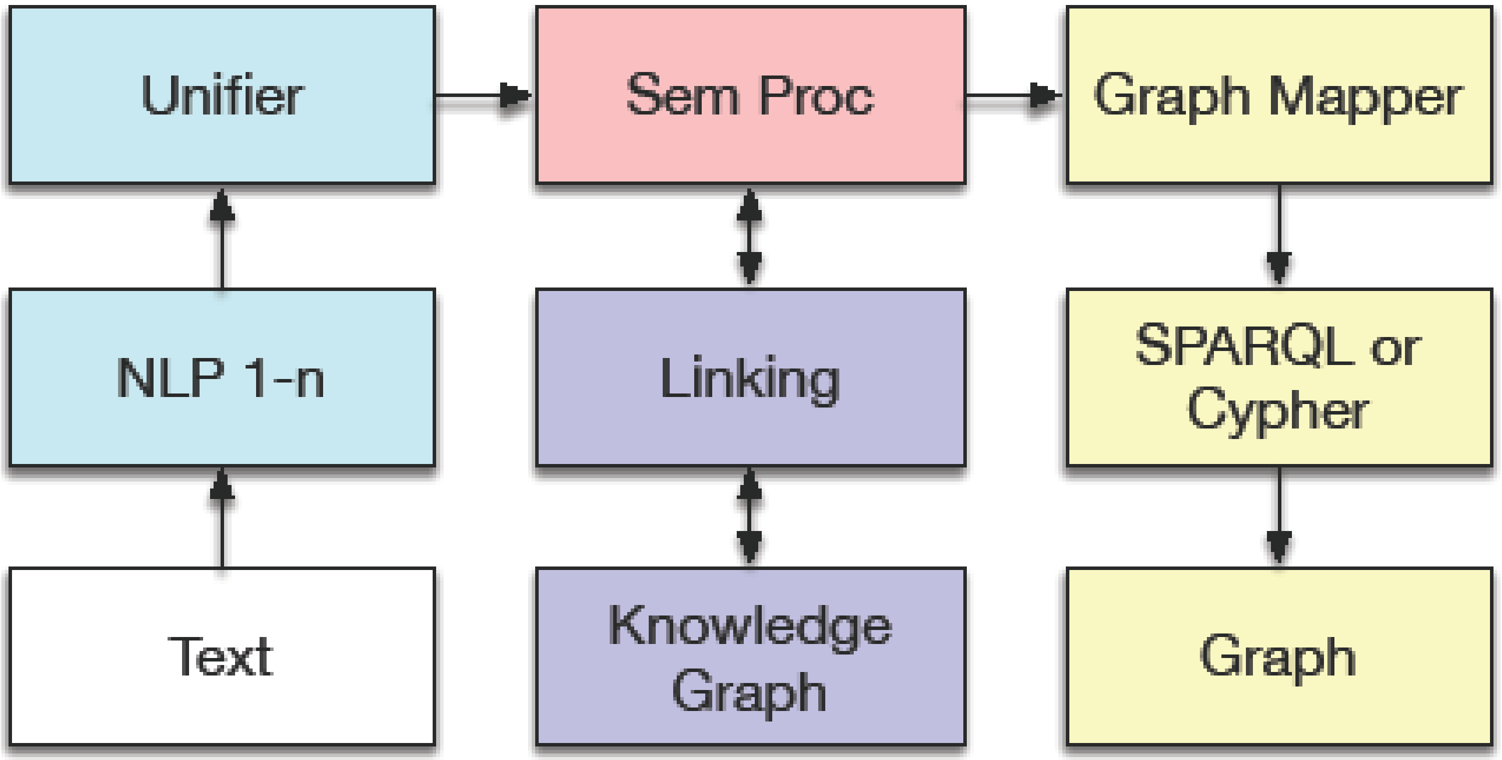


Knowledge Representations

- Practical use cases:
  - Dialogs
    - Topic and concepts in focus (conversational example)
  - Common Sense
    - Anaphora resolution using semantic properties
      - "Take the knife, cut the lime into two halves, and squeeze it." (p.c. Matthias Scheutz)
  - …

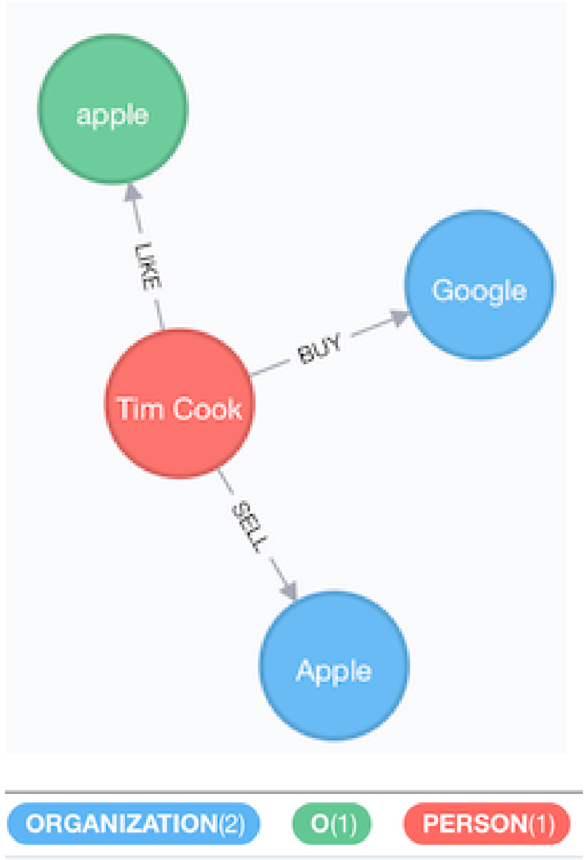
Pipeline

- Knowledge Graph Generation



Concept Relation Mapping

- Input:
  - Tim Cook sold Apple.
  - He bought Google.
  - He likes apples.
- 1st level typing using:
  - Named Entity Recognition



Linking

- Identification of the unique entity in a large Knowledge Graph
  - E.g. YAGO, DBpedia, ConceptNet, …
- Our approach:
  - Disambiguation using word and graph embeddings
- Language Independent
  - Language agnostic entity extraction

Typing

- Identification of the closest Hypernym
  - WordNet lookup
  - Microsoft Concept Graph
  - Using Linking results

Word and Graph Embeddings

- Distributional Semantics approach
  - Words are represented by vectors of a fixed length
  - Vectors are prediction models (e.g. Word2Vec):
    - Maximize the predicted likelihood of the words in their context

- Graph embeddings:
  - Semantic and conceptual: concepts and relations in graph context
  - Topological: shape of a conceptual sub-graph

Knowledge Representations

- General World Knowledge
  - From static to dynamic, with inferencing, reasoning
- Domain Specific Knowledge
  - Medical, Financial, Business, Legal, etc.
- Discourse specific Knowledge
  - Simple dialog memory (concepts and their linguistic features, relevant for anaphora resolution, coreference analysis)
  - Knowledge Graph or Ontology of semantic concept space in encapsulated discourse

Speech Acts, Implicatures, Presuppositions

- Deep Linguistic Processing:
  - A to B: “I bought the blue car.”
  - Implicature:
    - A and B talked about the event earlier.
    - There is a set of cars, at least 2 that was in the range of A’s action.
    - None of the other cars in the set is blue.
  - Linguistic indicators:
    - Definiteness via “the”
    - Specificity of the Noun Phrase

Speech Acts, Implicatures, Presuppositions

- Deep Linguistic Processing:
  - "Peter fed his cat."
  - Presupposition:
    - Peter owns a cat.
    - Peter owns cat food.
    - …
  - Linguistic indicators:
    - Possessive
- Types:
  - Universal linguistic properties (see Grice Maxims, Relevance Theory)
  - Language specific properties (dependency to cultural and sociological aspects)
  - Domain specific: e.g. "to be like milk"

HooSIER IE Approach

- Advanced NLP technologies
  - Deep linguistic processing
    - Tense, Voice, Mood detection
    - Hierarchical relations of elements in the clause, clause detection, scope reconstruction
  - Identification of phrasal heads of arguments, compound structure, and modifiers
  - Normalization of words and phrases
  - Extraction of core semantic relations
  - Extraction of modifiers and meta-information
  - Mapping of relations into complex Graphs (towards Description Logic representations)
  - Linking of entities and relations to Knowledge Graphs and Ontologies

HooSIER IE Approach

- Deep Linguistics
  - Tense, Voice, Mood detection
    - Tim Cook left Apple.
    - Tim Cook will leave Apple.
    - Apple was bought by Google.
  - Scope relations
    - Tim Cook did not leave Apple.
    - Tim Cook left, not Apple, but the board of Alphabet Inc.
  - Clause detection and scope
    - I wish [Tim Cook left Apple]
    - I did not claim [that Tim Cook left Apple]

HooSIER IE Approach

- Identification of phrasal heads of arguments, compound structure, and modifiers
  - The former president of the United States, Barak Obama…
  - Head: Obama
  - Compound component: Barak
  - Modification or Specification: “the former president of the United States”
- Mapping into complex Graphs
  - Concepts or entities
  - Relations between entities
  - Attribute-value pairs associated with entities and relations

HooSIER IE Approach

- Normalization of words and phrases
  - Lemmatization
    - *chatting*, *went*, *hired* → *chat*, *go*, *hire*
  - Reduction to core properties (semantic normalization)
    - X *was chatting with* Y → X – *talk* – Y
  - Multi-lingual normalization:
    - Machine translation prior to extraction of entity-relation tuples
    - Linking of entities and relations to a language neutral representation
      - More later (using YAGO, MS Concept Graph, VerbNet, PropBank etc.)

HooSIER IE Approach

- Extraction of core semantic relations
  - Predicate argument structures:
    - Tim Cook left Apple.
      - Predicate: leave
      - Argument 1 (subject, agent): Tim Cook
      - Argument 2 (object, patient or beneficiary): Apple
    - Tim Cook, who lives in San Francisco, left yesterday suddenly Apple without further explanation.
- Extraction of modifiers
  - Tim Cook – livesIn – SF
- Extraction of time references:
  - One day before document production time

HooSIER IE Approach

- Entities and relations as Graphs
  - Entities
    - String representation
    - Label = type
    - All other information:
      - Attribute-Value tuples associated with entity
  - Relations
    - String representation
    - Label – predicate type (e.g. PropBank ID)
    - All other information:
      - Attribute-Value tuples associated with entity
    - Relations have directionality, domain, and range
    - Domain and Range can be entities (and relations in some Graph DBs)

HooSIER IE Approach

- Linking of entities and relations to Knowledge Graphs and Ontologies
  - Large Knowledge Graphs as Link targets
    - Language independent URI/specification
    - Detailed concept properties
    - Multi-lingual representations or realizations of concept names
    - Example: DBpedia, YAGO, MS Concept Graph, Google KG, etc.
  - Ontologies (domain specific models, taxonomies)
    - Core taxonomy relations: isA hierarchy essential for efficient reasoning
    - Semantic type and consistency checking with assertions into graphs
    - Reasoning
  - Identification of the most specific hypernym for any entity/concept
    - THING – … – MAMMAL – DOG – POODLE
    - THING – … – FRUIT – APPLE
    - apple isA fruit
    - poodle isA dog

HooSIER IE Approach

- Typing of entities:
  - NLP-based pre-typing
    - Named Entity Recognition (NER) types: PERSON, ORGANIZATION, PLACE, DATE, TIME, CURRENCY, TITLE, … (5 to 7 core types of onomastic entities)
  - Knowledge Graph based typing
    - YAGO more than 17,000 types
  - Domain specific NER or Taxonomy-based typing
    - Our own model of types and potentially sub- or co-types
    - Develop own NER components
      - (Weighted) Finite State Transducers for (multi-) word analysis
      - Trained NER models using own corpora and data-sets

HooSIER IE Approach

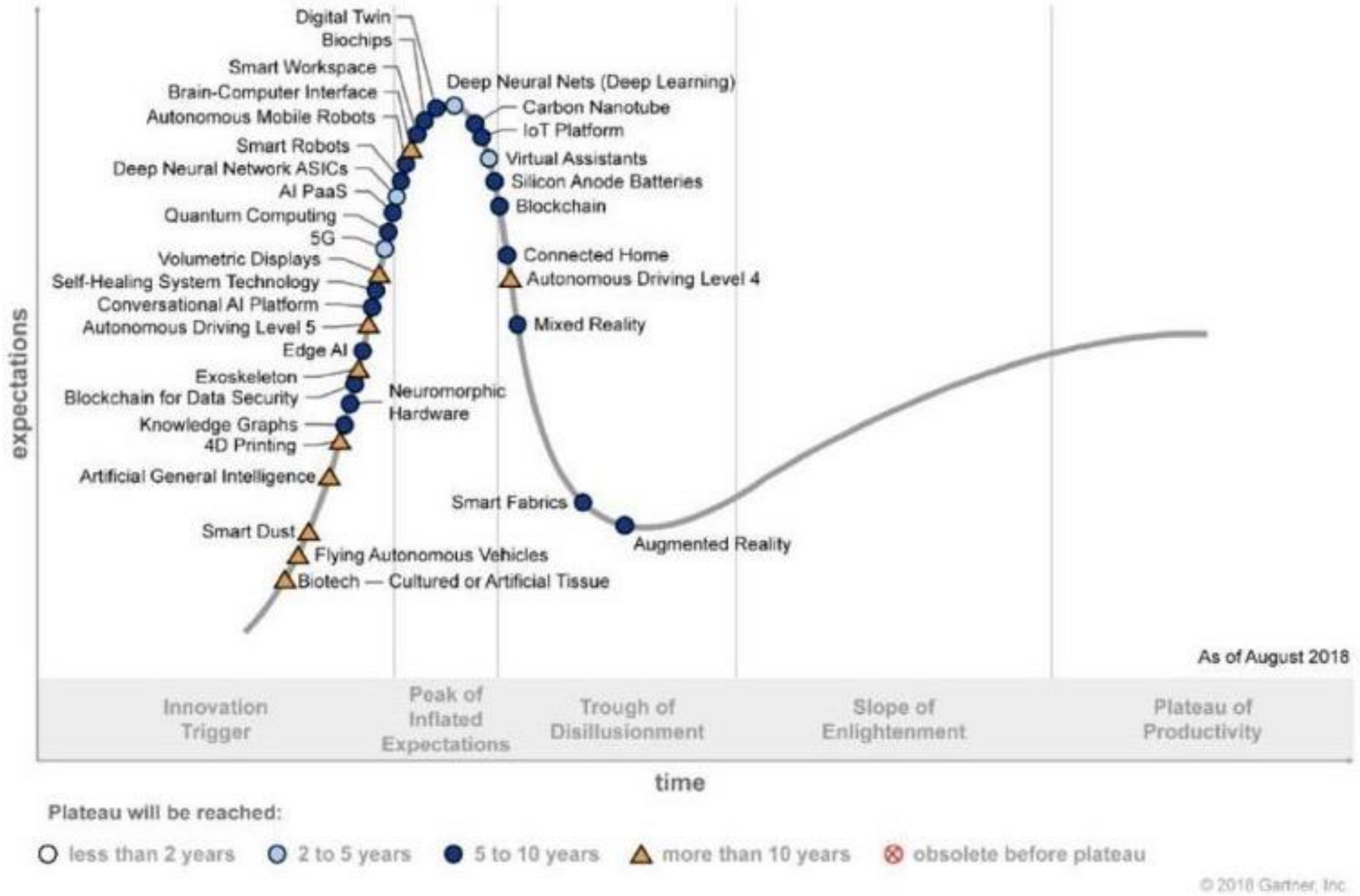
- Linking Disambiguation
  - Multiple types (hypernyms) for an entity in a given KG
  - NER types reduce the ambiguity
    - NLP components introduce error with NER
  - Use word embeddings and vector based models for disambiguation
    - Using Google, FastText, or GloVe vectors
    - Given vector for the target entity word (or multi-word expression) X
      - Tim Cook like apples. → X = apples/apple
    - For every hypernym candidate (and its hypernym, synonyms, and hyponyms) Y compute the probability of the observed context
    - Pick the one hypernym (and its semantic context) that best predicts the context of X

HooSIER IE Approach

- Expand Graph Representations (multiple graphs or linked sub-graphs)
  - Propositions represented as multiple entity-relation graphs
    - True propositions
    - Projected future related propositions
    - Assumed false propositions
  - Graph representation of Implicatures and Presuppositions
  - Entity identification and typing
    - Detailed semantic properties
    - Most specific type from isA taxonomy
    - Induction of types from Edge2Vec, predicate argument structures (e.g. VerbNet, PropBank), Graph similarity etc.
      - Syntagmatic vs. Paradigmatic relations

HooSIER IE Approach

- Applications:
  - Event identification and extraction (types: political event, pandemic outbreaks, civil unrest, security related events, etc.)
    - Agents, locations, time, timeline, causalities, victims, etc.
  - Graph-similarity as document similarity
  - Summarization using graph-based text generation
  - Search and query
    - Graph-search, e.g. query to graph and similarity search, graph navigation
  - Ontology or Knowledge Graph generation
    - Forensic, investigative
    - AI or chatbot related



Digital Twin
Biochips
Smart Workspace
Brain-Computer Interface
Autonomous Mobile Robots
Smart Robots
Deep Neural Network ASICs
AI PaaS
Quantum Computing
5G
Volumetric Displays
Self-Healing System Technology
Conversational AI Platform
Autonomous Driving Level 5
Edge AI
Exoskeleton
Blockchain for Data Security
Knowledge Graphs
4D Printing
Artificial General Intelligence
Smart Dust
Flying Autonomous Vehicles
Biotech — Cultured or Artificial Tissue
Neuromorphic Hardware
Deep Neural Nets (Deep Learning)
Carbon Nanotube
IoT Platform
Virtual Assistants
Silicon Anode Batteries
Blockchain
Connected Home
Autonomous Driving Level 4
Mixed Reality
Smart Fabrics
Augmented Reality
expectations
As of August 2018
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
© 2018 Gartner, Inc.

Technologies

- Environment
  - Microservices using isolated RESTful modules
  - Mainly Java, Scala, Apache Spark
    - Wrapping C(++), Python
  - Databases
    - MongoDB, PostgreSQL hosting Knowledge Graphs (DBpedia, YAGO, MS Concept Graph)
  - Neo4J (Cypher), Stardog (SPARQL & OWL)
  - Docker Containers

Thanks

- NLP-Lab students:
  - Oren Baldinger, Maanvitha Gongalla, Yiwen Zhang, Umang Mehta, Shreejith Panicker, Aarnav, Abhishek Babuji, Richard Xu, Chaitanya Patil, Gopal Seshadri, Shujun Liu, Peace Han, Rohit Bapat, Anurag Kumar, Murali Kishore, Varma Kammili, Mahesh Latnekar, Aarushi Bisht, Jagpreet Singh Chawla, …
- Lwin Moe