PA164 Natural Language Learning

Lecture 12: Machine Learning for Knowledge Extraction from Text

Vít Nováček

Faculty of Informatics, Masaryk University

Autumn, 2021

MUNI

Outline

What Is Knowledge, Actually?

- Well, who knows. . .
- But the approximate consensus (based on Oxford dictionary) is more or less this:
  - facts, information, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject
  - what is known in a particular field or in total; facts and information
  - certain understanding, as opposed to opinion
  - awareness or familiarity gained by experience of a fact or situation
- Knowledge representation
  - Computer science discipline (a specific part of AI)
  - Dealing mostly with knowledge that can be formalised via logics
  - Other (more practical) approaches gaining prominence recently, though

What Is Knowledge Extraction, Then?

- Creation of knowledge from data that can be
  - structured (e.g. relational databases, XML or HTML), or
  - unstructured (e.g., text, speech, images or video)
- Conceptually related to NLP or ETL
- Typically, however, knowledge extraction assumes:
  - either reusing of formal knowledge (a machine-readable vocabulary or an ontology), or
  - induction of some sort of formal schema from the data

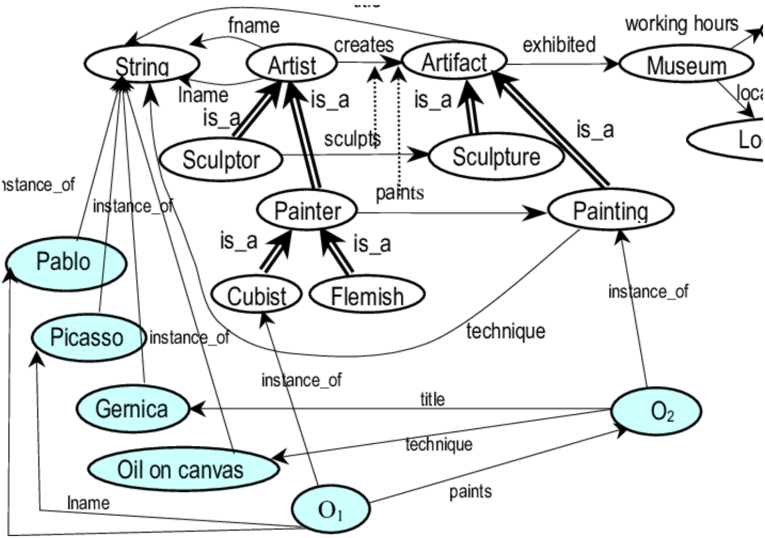
Example of a KR Formalism—Ontologies

- Representation, formal naming and definition of general categories (also called concepts or classes) and individuals falling under them
- Properties of the categories and individual entities, relationships between them
- Metadata and annotations that do not affect the formal meaning
- Typically based on subsets of first order predicate logic called Description Logics
- Allow for deductive reasoning (typically)
- Sophisticated, but pretty heavy-weight and expensive to create and maintain

Example of a KR Formalism—Knowledge Graphs

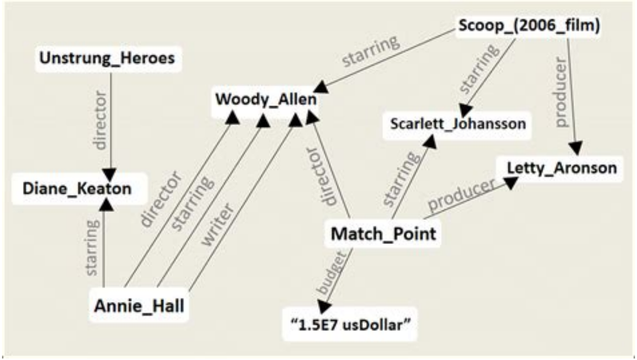
- Still formal, but more relaxed knowledge representation
- Based on linked representation of data in the form of subject-predicate-object triples
- Much more flexible and easier to maintain
- Amenable to inductive, and, more recently, also transductive reasoning
- Inference (by learning) can benefit from recent advances in neural information processing

An Ontology Example



[1] Danger, Roxana, et al. "A proposal for the automatic generation of instances from unstructured text." Iberoamerican Congress on Pattern Recognition. Springer, Berlin, Heidelberg, 2004.

A Knowledge Graph Example



[2] Arnaout, Hiba, and Shady Elbassuoni. "Effective searching of RDF knowledge graphs." Journal of Web Semantics 48 (2018): 66-84.

Knowledge Extraction by Means of Ontology Learning

- Automated or semi-automated process of knowledge extraction
- Typically from text or semi-structured resources (such as Wikipedia)
- The output is a variously complex ontology (or a knowledge graph)
- Can consist of refinement or population of an existing ontology
- Leverages many computational linguistics and machine learning techniques

Why Bother?

- Creating machine-readable knowledge bases manually is expensive and error-prone
- Yet they are useful for plenty practical things
  - Development of intelligent software agents (e.g., chatbots), question answering apps
  - Robotics
  - Quality features for machine learning algorithms
  - Ground truth and background knowledge for hybrid machine learning techniques
  - Knowledge bases for explainable AI
  - ...
- High degree of automation of the process is thus very desirable
  - Deals (to some extent) with human bias in creating knowledge
  - Is way more scalable and less expensive
  - Can often be relatively easily ported between different domains

Typical Ontology Construction/Learning Tasks

- Term extraction
  - A prerequisite for all aspects of ontology construction or learning—basic units of meaning (words, phrases)
- Synonym discovery
  - Aims to find the terms that indicate the same concept
- Concept formation
  - A formal representation of the concept intention, extension and the lexical signs (terms) which are used to refer to it
  - Rather blurry and contested task, though
- Establishing concept hierarchy
  - Build the hierarchical taxonomy of concepts (hypero-hyponymy relations)
- Relation discovery (or extraction)
  - Extracting novel relationships between known concepts
- Rule or axiom extraction
  - A pinnacle of ontology construction—inferring logical rules and axioms based on extracted concepts and relations

Main Challenges to Ontology Learning

- Noisy, dynamic and large input data
- Sparse and/or unbalanced labelled data
- Lack of consensus on some basic definitions (and resulting difficulties in defining the problems to be solved formally enough)
- Lack of validation resources
- Under-researched quantitative evaluation methodologies
  - Precision and recall often used as proxies
  - Rather coarse-grained, though
  - Alternative metrics may be too qualitative

Overview of Approaches to Ontology Construction

- Manual approaches (ontology engineering)
- (Semi)automated approaches
  - Linguistics-based
  - Logics-based
  - Classical machine learning
  - Deep learning
  - Hybrid approaches

Outline

Methods Based on Normative Linguistics

- Pattern-based extraction
  - Recognizing relations by matching patterns against word sequences
  - Employs lexico-syntactic patterns and semantic templates (e.g., “NP is type of NP” for hypernyms)
  - Reasonable precision, but very low recall
- POS tagging and sentence parsing
  - Essentially a rule-based approach
  - POS tagging to categorise words in the text, parsing to recover context to disambiguate
  - Mostly used for term extraction
- Syntactic and dependency structure analysis
  - Utilising sentence structure and dependencies to extract
    - terms (e.g., complex noun phrases), and
    - relationships (subject-predicate-object triples derived from the corresponding syntactic elements)

Methods Based on Statistics

- Co-occurrence analysis
  - Finding lexical units that tend to occur together
  - Used for anything between term extraction and discovering implicit relations between concepts
- Association rules
  - Extracting non-taxonomic relations between concepts
  - Typically, using a small seed knowledge as background (e.g., a taxonomy)
- Heuristic and conceptual clustering
  - Grouping concepts based on the semantic distance between them to make up hierarchies
  - Formal Concept Analysis (FCA) as a possible method
    - Conceptual clustering based on lattices and ordered sets to provide intentional descriptions for concepts
- Ontology pruning
  - Building a domain relevant ontology by using heterogeneous sources
  - E.g. comparing domain sources with generic sources...
  - to determine which concepts are more relevant to the specific domain and which concepts are general

Methods Based on Logics and Inference

- Inductive Logic Programming
  - Deriving rules from positive and negative examples of the existing collection of concepts
  - E.g., “cats have fur”, “dogs have fur”, “tigers have fur” $\longrightarrow$ “mammals have fur”
  - Continuous refinement of the rules based on further examples (e.g., “humans don’t have fur”)
- Logical inference
  - Deriving implicit knowledge by means of deductive reasoning via seed facts, axioms and inference rules
  - Tends to generate obvious relations, and/or suffer from inconsistencies in real-world data

Outline

Term Extraction, Synonym Discovery and Concept Formation

- Often, the afore-mentioned NLP techniques are used for this
- The results are then used for bootstrapping the consequent layers
- An example:
  - Get terms by POS tagging and parsing
  - Learn concepts (groups of terms) and their taxonomy by means of hierarchical clustering

Taxonomy Extraction

- Classical unsupervised clustering
  - Pretty much any standard algorithm can be used
  - Typically makes use of vector space representation of the textual data
  - Can employ word embeddings, too
- Formal Concept Analysis (FCA)
  - Based on mathematical order theory
  - Formalisation of concept extension and intension
  - A formal concept is defined to be a pair (A, B)...
  - where A is a set of objects (called the extent), and...
  - B is a set of attributes (the intent) such that:
    - the extent A consists of all objects that share the attributes in B, and, dually,
    - the intent B consists of all attributes shared by the objects in A.
  - Formal concepts can then be ordered in a hierarchy ("concept lattice")

Relation Extraction

- Pattern-based techniques heavily used
  - Define seed lexico-semantic patterns for a relation
  - Bootstrap more patterns automatically based on context in a corpus
  - Discover relations by pattern-matching in the text
- Conditional Random Fields for extracting concept attributes
- Named entity recognition followed by defining a dataset of seed relations and clustering these with unseen texts
- Still a rather under-researched field, though

Rule or Axiom Extraction

- Even more experimental than relation extraction
  - Some rule-based techniques again, defining axiom templates
  - Dependency parsing trees can also be used
  - Semantic similarity and association rule mining for generating disjointness relations
- Most techniques dependent on the (often dubious) quality of the previous steps, though

Outline

Prominent Ontology Learning Tools (1/2)

System Name	Input Type & Language	Learned Elements: Terms	Concepts	Taxonomic Relations	Non-taxonomic Relations	Axioms	Used Approach	User Intervention	Output Format	P & R, or F %
ASIUM	Unstructured text French	√					Syntactic structure analysis	Whole process	Frame based	Not provided
			√							
				√			Conceptual clustering			
CRCTOL	Unstructured (plain text documents only) English	√					Syntactic structure analysis, POS tagging, & relevance measures	Validation & Evaluation	RDFS or OWL	Not provided
			√							
				√			Lexico-syntactic patterns, syntactic structure analysis			
					√					
DODDLEII	Unstructured & structured data English			√			co-occurrence (4-grams) & association rules algorithm	Whole process	Information not provided	P= 23, R= 56
					√					
HASTI	Unstructured Persain	√					Lexico-syntactic patterns & semantic templates	Validation & Evaluation	Subset of KIF	Not provided
			√							
				√			Semantic templates, heuristic clustering & logical inference			
					√					
						√	Inductive logic programming			
OntoCmaps	Unstructured English	√					Dependency structure analysis & POS tagging	Validation & Evaluation	OWL	Not provided
				√			Dependency structure analysis, hierarchical clustering & filtering matrices			
					√					
SYNDIKATE	Unstructured text German	√					Syntactic structure analysis & Dependency structure analysis	Evaluation	Special format	[1] (P= 97, R= 57)* (P= 94, R= 31)**
			√							
				√			Dependency structure analysis & Semantic templates			
					√					

[1] Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. "Automatic ontology construction from text: a review from shallow to deep learning trend." Artificial Intelligence Review 53.6 (2020): 3901-3928.

Prominent Ontology Learning Tools (2/2)

System Name	Input Type & Language	Term s	Concepts	Taxonomic Relations	Non-taxonomic Relations	Axioms	Used Approach	User Intervention	Output Format	P & R, or F %
		Learned Elements								
TEXT2O NTO	Unstructured text Spanish & German	√					POS tagging, Syntactic structure analysis& relevance metrics	Validation & Evaluation	F-Logic, OWL or RDFS	F=22, P= 17, R= 30
			√				formal concept analysis			
				√			hierarchical clustering & lexico-syntactic patterns			
					√		association rules			
TextStor m and Clouds	Unstructured English	√					POS tagging, &Syntactic structure analysis	Whole process	Part of Dr. Divago project	F=52
				√						
					√					
						√	Inductive logic programming			
TEXT-TO-ONTO	Structured, or sim-structured German	√					POS tagging, & Syntactic structure analysis	Validation & Evaluation	F-Logic, RDFS / DAML+OIL, &Part of KAON	Not provided
			√				Formal concept analysis & pruning			
				√			Hierarchical clustering & lexico-syntactic patterns			
					√		Association rules			
PROMIN E	Sim-structured English	√					POS tagging & relevance measures	Validation & Evaluation	Subset of PROKEX project	P= 89, R= 86
			√							
				√			Heuristic clustering& filtering measures			

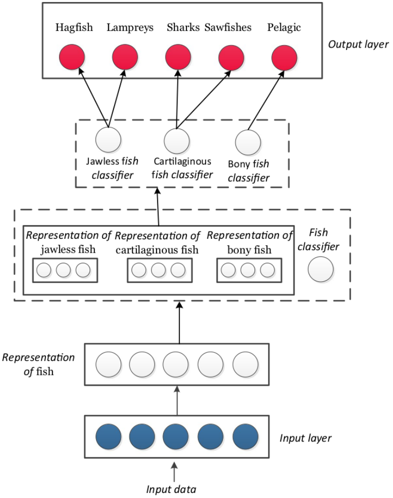
[1] Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. "Automatic ontology construction from text: a review from shallow to deep learning trend." Artificial Intelligence Review 53.6 (2020): 3901-3928.

Outline

Rationale of the Deep Learning Approaches

- Motivated by the problem of limited language understanding by machine using shallow processing for text
- There's a hope that the representation learning aspect of DL approaches could help
- No full-fledged DL framework for ontology learning mature enough yet
- Some promising approaches exist already, though, such as deep learning models for
  - extracting entity attributes
  - extracting specific instances of pre-defined relationship types
  - named entity recognition
  - learning word embeddings, followed by taxonomy construction
  - transductive reasoning for concerting natural language into a formal one (OWL)
  - semi-automated ontology construction based on text classification and TF-IDF scoring
  - autoencoders for enriching Gene Ontology by newly discovered gene functions
  - ...

Example—Deep Learning for Concept Classification



[1] Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. "Automatic ontology construction from text: a review from shallow to deep learning trend." Artificial Intelligence Review 53.6 (2020): 3901-3928.

Selected Deep Learning Approaches (1/2)

Task	Study	DL model	Language	Input	Domain	Target	Other Details
Term extraction and relation discovery	Albukhitan et al. (2017)	CBOW and Skip-gram	Arabic	5 thousand words	Not determined	5022 concepts and 830 relations	The system extracted correctly 3861 concepts and 587 relations
Axiom learning	Arguello Casteleiro et al. (2017)	CBOW and Skip-gram	Not mentioned	301,202 PubMed publications (title and abstract)	Biomedical (sepsis)	Get the candidate terms related to sepsis	
Relation discovery	Chen et al. (2010)	DBN	Chinese	221 documents	5 entity types (Person, Organization, GPE, location, and facility	5 types of relations (Role, Part, At, Near, and Social)	Dataset is ACE 2004
Term extraction and relation discovery	Chicco et al. (2014)	Autoencoder	Not mentioned	Bos taurus (cattle) and Gallus gallus, (red junglefowl), gene sets from the Genomic and proteomic data warehouse (GPDW 2009 and 2013)	Biomedical (Gene)	Create and enrich gene database with massive gene function annotation and prediction	
Term extraction and relation discovery	Hassan and Mahmood (2018)	CNN and RNN	Not mentioned	Stanford Large Movie Review dataset (IMDB) and the Stanford Sentiment Treebank dataset (SSTb)	Sentiment analysis	Sentence classification	8544 sentences for training, 2210 for testing, and 1101 for validation

[1] Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. "Automatic ontology construction from text: a review from shallow to deep learning trend." Artificial Intelligence Review 53.6 (2020): 3901-3928.

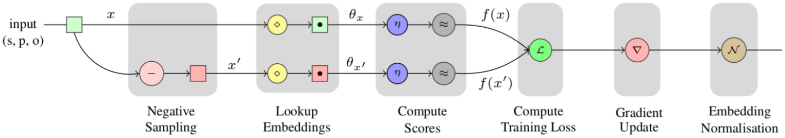
Selected Deep Learning Approaches (2/2)

Task	Study	DL model	Language	Input	Domain	Target	Other Details
Axiom learning	Neelakantan (2017)	RNN	English	Freebase and Google's entity linking in ClueWeb, with 10 k entity pairs and 2 million paths per relation type	Several domains (not determined)	to produce latent programs involving arithmetic and logic operations	Use WikiTableQuestions dataset as test set
Axiom learning	Petrucci et al. (2016)	RNN	English	123 millions sentences and formulas	Several domains (not determined)	learning expressive ALCQ axioms	Dataset collected from encyclopedias and previous studies
Relation discovery	Wang (2015)	DBN	Not mentioned	Not determined	Different domains (not determined)	3 types of relations (subclass, disjoint, and coexists)	Use top-down DBN
Relation discovery	Wang et al. (2018)	CNN	Chinese	68,000 texts	Shipping industry	classify the terminology of domain category	47,600 texts for training set, 13,600 for verification set and 6800 for test set
Relation discovery	Zhong et al. (2016)	CRF and DBN	Chinese	24 MB	Crawler in the shipping news and travel websites	4 categories of entities attributes (Port, ship, routes, and view)	more than 10,000 sentences as training set 26 extracted attributes

[1] Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. "Automatic ontology construction from text: a review from shallow to deep learning trend." Artificial Intelligence Review 53.6 (2020): 3901-3928.

KGEs for Relation Extraction (1/2)

- Knowledge graph embeddings (KGEs):
  - A supervised machine learning problem
  - Falls under statistical relational learning
  - Effectively, fitting a multivariate probability density function...
  - to the positive and negative "links" (i.e. subject-predicate-object triples) in a knowledge graph



KGEs for Relation Extraction (2/2)

- An example of a method for relation extraction by means of KGEs:
  - The plausibility of each missing fact $< s, p, o >$ in the KG can be predicted as $score(< s, p, o >)$
  - A text-based model can be used to similarly score the similarity between each relation p and its textual mention in an input corpus
  - These scores can then be combined to train a joint text-KG embedding model
  - This model refines the predictions of extracted relations based purely on the text
- A several other, slightly different approaches have been proposed, too

Outline

Further Readings (1/2)

- Ontologies and knowledge graphs in general:
  - Staab, Steffen, and Rudi Studer, eds. "Handbook on ontologies." Springer Science & Business Media, 2010.
  - Hogan, Aidan, et al. "Knowledge graphs." Synthesis Lectures on Data, Semantics, and Knowledge 12.2 (2021): 1-257.
- Recent survey on ontology learning:
  - Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. "Automatic ontology construction from text: a review from shallow to deep learning trend." Artificial Intelligence Review 53.6 (2020): 3901-3928.
- Ontology learning classics:
  - Maedche, Alexander, and Steffen Staab. "Ontology learning for the semantic web." IEEE Intelligent systems 16.2 (2001): 72-79.
  - Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini, eds. Ontology learning from text: methods, evaluation and applications. Vol. 123. IOS press, 2005.
  - Asim, Muhammad Nabeel, et al. "A survey of ontology learning techniques and applications." Database 2018 (2018).

Further Readings (2/2)

- Approaches based on knowledge graph embeddings:
  - Wang, Quan, et al. "Knowledge graph embedding: A survey of approaches and applications." IEEE Transactions on Knowledge and Data Engineering 29.12 (2017): 2724-2743.
  - Wang, Zhen, et al. "Knowledge graph and text jointly embedding." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.