What I learned about Research in the Digital Age

Sanda Harabagiu

University of Texas at Dallas



Why Research ????

- Because I always was *curious* and *enjoyed learning*
- Because being creative and innovative is akin to being an ARTIST of knowledge
- Research allows you to practice your art of knowledge !!!













- How could a system automatically find the answer to any question?
- How could you build a system?
- How could you represent knowledge?
- What is intelligence?
- How do humans communicate?
- Why we understand each other?
- Why is natural language difficult?
- What did I learn about semantics after building successful Q/A systems?



Textual Question Answering





Problem complexity





Semantics is just one of the problems! But what kind of semantics do we need?





Semantic Problems with EATs

- The problem of assigning EATs
 - E.g. "manner questions":
 - Example "How did Hitler die?"



- The problem of recognizing answer types/structures
 - Should "manner of death" be considered an answer type?
 - What other manner of event/action should be considered as answer types?
- The problem of recognizing EATs in texts
 - Should we learn to extract "manner" relations?
 - What other types of relations should we consider?
 - Is relation recognition sufficient for answering all types of questions? Is it necessary?





EAT = Manner-of-death



In TREC evaluations several questions asked about manner of death:

- "How did Adolf Hitler die?"
- <u>Solution:</u>
 - We considered "Manner-of-Death" as an answer type, pointing to a variety of verbs and nominalizations encoded in WordNet
 - We developed text mining techniques for identifying such information based on lexico-semantic patterns from WordNet
 - Example:
 - [kill #sense1 (verb) CAUSE \rightarrow die #sense1 (verb)]
 - Source of the troponyms of the [*kill #sense1 (verb)*] concept are candidates for the MANNER-OF-DEATH hierarchy
 - e.g., drown, poison, strangle, assassinate, shoot

Practical Hurdle



Lesson #3

Semantic information for EATs needs to be recognized by text mining techniques

- Not all MANNER-OF-DEATH concepts are lexicalized as verbs

 → we set out to determine additional patterns that capture such cases
- <u>Goal:</u> (1) set of patterns
 - (2) dictionaries corresponding to such patterns
 - \rightarrow well known IE technique: (IJCAI'99, Riloff&Jones)

$ \left[\begin{array}{c} X \left\{ DIE \\ be \text{ killed} \end{array} \right\} \text{ in ACCIDENT} \right] $	seed: train, accident, (ACCIDENT) car wreck
X DIE {from of} DISEASE be killed	seed: cancer (DISEASE) AIDS
X DIE after suffering MED Suffering of CON	ICAL seed: stroke, IDITION (ACCIDENT) complications caused by diabetes

Results: more than100 patterns were discovered





By doing only:

- named entity recognition
- semantic classification of the expected answer type (off-line taxonomy + semantic info from WordNet)
- Text mining

55% accuracy on factual trivia-like questions (TREC-8, Moldovan et al.)

What else is needed?

Most questions cannot be processed successfully in this way, as they do not have a simple, conceptual EAT We need to consider additional forms of semantic knowledge

and semantic processing.





AQUINAS – Answering QUestions using INference and Advanced Semantics



The driving rationale for our approach is that humans appear to have limited need for factoid question answering, but rather much more need to have systems that can deal with <u>complex reasoning about</u> causes, effects and chains of hypotheses.



QA architecture based on semantic structures





Applying Predicate-Argument Structures to QA

Predicate-arguments structures improve answer extraction!!!

Parsing Questions

Q: What kind of materials were stolen from the Russian navy?

PAS(Q): What [<u>Arg1</u>: kind of nuclear materials] were [Predicate:<u>stolen</u>] [Arg2: from the Russian Navy]?

• Parsing Answers

A(Q): Russia's Pacific Fleet has also fallen prey to nuclear theft; in 1/96, approximately 7 kg of HEU was reportedly stolen from a naval base in Sovetskaya Gavan.

PAS(A(Q)): [Arg1(P1redicate 1): Russia's Pacific Fleet] has [ArgM-Dis(Predicate 1) also] [Predicate 1: fallen] [Arg1(Predicate 1): prey to nuclear theft]; [ArgM-TMP(Predicate 2): in 1/96], [<u>Arg1(Predicate 2): approximately 7 kg of HEU]</u> was [ArgM-ADV(Predicate 2) reportedly] [Predicate 2: <u>stolen</u>] [Arg2(Predicate 2): from a naval base] [Arg3(Predicate 2): in Sovetskawa Gavan]

• **Result: exact answer**= "approximately 7 kg of HEU"





Lesson #5

Additional semantic resources: FrameNet

- Using FrameNet for QA
- **<u>Example</u>**: What stimulated India's missile programs?

FRAME: Stimulate

Frame Element CIRCUMSTANCES: <u>ANSWER (part 1)</u> Frame Element EXPERIENCER: India's missile program Frame Element STIMULUS: <u>ANSWER (part 2)</u>

FRAME: Subject_Stimulus

Frame Element CIRCUMSTANCES: <u>ANSWER (part 3)</u> Frame Element COMPARISON SET: <u>ANSWER (part 4)</u> Frame Element EXPERIENCER: India's missile program Frame Element PARAMETER: nuclear proliferation





- Produced by ICSI Berkeley [Baker et al.,1998] as a lexico-semantic resource encoding a set of frames (schematic representations of situations)
- Frames are characterized by:
 - target words or lexical predicates whose meaning includes aspects of the frame;
 - frame elements (FEs) which represent the semantic roles of the frame;
 - examples of annotations performed on the British National Corpus for instances of each target word.
- The project methodology was done on a *frame-by-frame basis*:
 - choosing a semantic frame (e.g. Commerce)
 - define the frame and its frame elements (e.g. BUYER, GOODS, SELLER, MONEY)
 - list the various lexical predicates which invoke the frame (buy, sell)
 - finding example sentences of each predicate in a corpus





Shallow Semantic Parsing Based on FrameNet

- Cosmin Adrian Bejan, Alessandro Moschitti, Paul Morărescu, Gabriel Nicolae and Sanda Harabagiu
- University of Texas at Dallas
- Human Language Technology Research Institute





Capitalizing on Clinical Data





BIG DATA





Data Science Home / BD2K Home Page

Big Data to Knowledge (BD2K)

The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.

BD2K Recent News

BD2KCenters Coordination Center Solicits Proposals for BD2K-Related Hackathons

The Big Data to Knowledge Centers Coordination Center is announcing a call for hackathon... *read more*



AUTOMATIC DISCOVERY AND PROCESSING OF EEG COHORTS FROM CLINICAL RECORDS

This award supports software development for automated explanatory modeling of complex healthcare data. The researchers develop a patient cohort retrieval system to provide big mechanism modeling capability for analysis of electroencephalogram (EEG) data.

Big mechanisms have been defined as large explanatory models of complex systems with many causal interactions. The project is centered on the aggregation of clinical knowledge automatically discovered from EEG signals and EEG reports into a medical knowledge graph.

The software framework established by this project could be transformative for mining the wealth of biomedical knowledge available from hospital medical records.

Research supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468.





Patient Cohort Retrieval:

Given a query modeling of the desired attributes of a patient subpopulation, return a ranked list of patients that match this criteria.

In large EMRs:

- 95k de-identified clinical records resulting from 18k hospital visits
- **Example Query:** Patients diagnosed with localized prostate cancer and treated with robotic surgery.

In EEG Reports:

- 20k de-identified EEG reports
- Example Query: patients with occasional sharp waves suggesting a potential for seizures







Electro-Enchephalography EEG

Patient Cohort Retrieval System

Query:









Patient Cohort

Query Results:

Feature (ACTIVALIPO3) Feature (ACTIVALIPO3)	
CLINICAL HISTORY: This is a 12-year-old male with a learning disability and possible seizures.	CLINUTL REFORT . THE IS A TTYPE YOU ARE WELL ARE UNIT OFFICIAL.
2 MEDICATIONS: None.	Transment (MACHALIPOS) Transment (MACHALIPOS) Transment (MACHALIPOS) Transment (MACHALIPOS) MEDICATIONS: Neuroperatural Machalipos assente Extension and Extended assente Extended
INTRODUCTION Data Make FEG. is performed in the law using standard 10-20 system of electrode placement with one channel of FKG.	medications, reconsidered accurry gyrine, remaining, Location, and Propositi
RATING BATTING BAT	INTRODUCTION: Digital video EEG is performed in the OR using standard 10-20 system of electrode placement with one channel of EKG.
4 Hyperventilation and photic stimulation are performed.	The record begins at 10:17 AM and concludes at 12:14 PM.
5 This is an awake and briefly drowsy record.	DESCRIPTION OF THE RECORD: As the tracing begins, the patient is in the OR.
EEG Anri-ley (#Active	EE6 Activity (FACTINAL)(POS) There is a brief awake pattern, followed by the appearance of an anesthetic pattern at 10.24.
EEG Activity (FACTUALIPOS) EEG Activity (FACTUALIPOS) Evidential (FACTUALIPOS)	There is a bit more delta on the left compared to the right.
7 There is a modest amount of background theta and occasional LAMEDA waves are identified.	EEG Activity (FACTUAL(POS) EEG Activity (FACTUAL(POS) EEG Activity (FACTUAL(POS)
EEG Event (FACTUAL)(FOS) EEG Expression (FACTUAL)(FOS)	The overall anesthetic pattern is one rich in beta frequency activity with a subtle asymmetry in delta.
Hyperventilation Is performed with good effort producing bursts of rhythmic, symmetric theta.	response to graver as a a.e.
Eliferetel (Active) (Acti	EE Anning #Activat_pros A subble asymmetry with a bit more delta on the left is again discernable at that time.
EE Event (FACTUAL)[PO3] Activity (FACTUAL)[PO3]	EE0 Activity [PACTUAL][POS]
0 Photic stimulation elicits driving.	The left carotid is clamped at 11:26:35, the external at 11:27. 36.
	EEG Activity (FACTUAL (FOS)
1 HR 72 BPM.	The common carotid is reclamped at 11:28.
INTERCOLUER DATA SEC. In a shift of bits and do by	EG Astivity (FACTUALISEG) Problem (FACTUALISEG) Evic (FACTUALISEG)
2 INFRESSOR, MILITY AUTOMILE EEG TOTA CHILO OF DIS AGE GUE ID.	The caroling in this signify asymmetric patient seen with placement of the camps, and no accuracians changes are noted information in the caroling of the caroling and the caroling of the car
Beta Activity PACTURAL POST	
F54 Antoine #ACTIAL 19031	The record did not indicate that a doppler study was performed, and the recording concludes at 12:14.
4 Mild excess theta.	Res and a second s
EGACUMU FACTUALINEO] EVISIONAL FACTUALINEOS	HR. 60 BPM.
S CLINICAL CORRELATION: No epileptiform features were identified in this record.	Hobien FACTURALIPOS ESSENT FACTURALIPOS INVESTOR FACTURALIPOS REVENUE FACTURALIPOS
Roblem (FACUAL (POS)	IMPRESSION/CLINICAL CORRELATION: EEG monitoring during this carolid endarterectomy was remarkable for some asymmetry seen prior
.6[Correlation with history and imaging may be of use.	
	but without change with placement of the clamp and without change throughout the course of the study.
	These findings are similar to the patient's baseline at presentation one week ago.



Temporal Disease Modelling:

The medical signs, symptoms, diagnoses, treatments, tests, and observations associated with a patient change over time. Thus, automatically reasoning about patients (e.g. for patient cohorts) can be improved by modelling the temporal aspect of the clinical picture and therapy.

Knowledge Discovery & Representation



Digital Age





$\mathsf{BIG}\;\mathsf{DATA}\to\!\mathsf{BIG}\;\mathsf{KNOWLEDGE}\to\mathsf{BIG}\;\mathsf{MECHANISMS}$



Thank you!!





