Ontology Based Legal Information Extraction

Group 27
Final Year Project - Mid Evaluation
University of Moratuwa

Introduction

1. Information Extraction
2. Ontology
3. Ontology Based Information Extraction

Information..

- Information Retrieval
- Information Extraction
- Natural Language
   Understanding

Information Extraction

Aims to process natural language text and to retrieve occurrences of a particular class of objects or events, and occurrences of relationships among them [1].

- Instances of street addresses from web pages
- Instances of storms from weather reports
- Instances of directors of movies
- Instances of co-authors of papers
- Instances of coach-player relationships

[1]. S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, pages 848-850, 2003

"The Social Network" remains at the top of the box for a second week grossing $15.5 million dollars. "The Social Network" is based on the life of founder Mark Zuckerburg, and co-founder Eduardo Saverin forming the social network of our generation "Facebook".

The movie starts with Mark and Eduardo's friendship ending in a nasty lawsuit. The two founders recollect on the creation of the website, and the series of events that led them to the lawsuit over power and money. The movies slogan '*You can't make 500 million friends without a few enemies*' is proven throughout the trials set before the once loyal companions.

"The Social Network" has a supporting role played by Justin Timberlake who takes on the character Sean Parker the creator of Napster. Parker assists Zuckerburg in the upbringing of Facebook, and the downfall between business partners. This biographical film reveals suspense and heavy case drama for the lives of these Harvard dropouts making and breaking bonds on the journey to success. This is a striking film with an underlying message of true moral. Overall, "The Social Network" is definitely worth going to see.

Movie: The Social Network
Actor: Justin Timberlake
Character: Mark Zuckerburg
Character: Eduardo Saverin
Character: Sean Parker

Ontology

A formal and explicit specification of a shared conceptualization[2, 3].

An Ontology has;

- Classes
- Data type properties
- Object properties (including taxonomical relationships)
- Instances (objects)
- Property values of the instances
- Constraints

[2] T. Gruber. Ontolingua: A Translation Approach to Providing Portable Ontology Specifications. Knowledge Acquisition, 5(2):199–220, 1993.
[3] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. Data Knowledge Engineering, 25(1-2):161–197, 1998.

Example Ontology
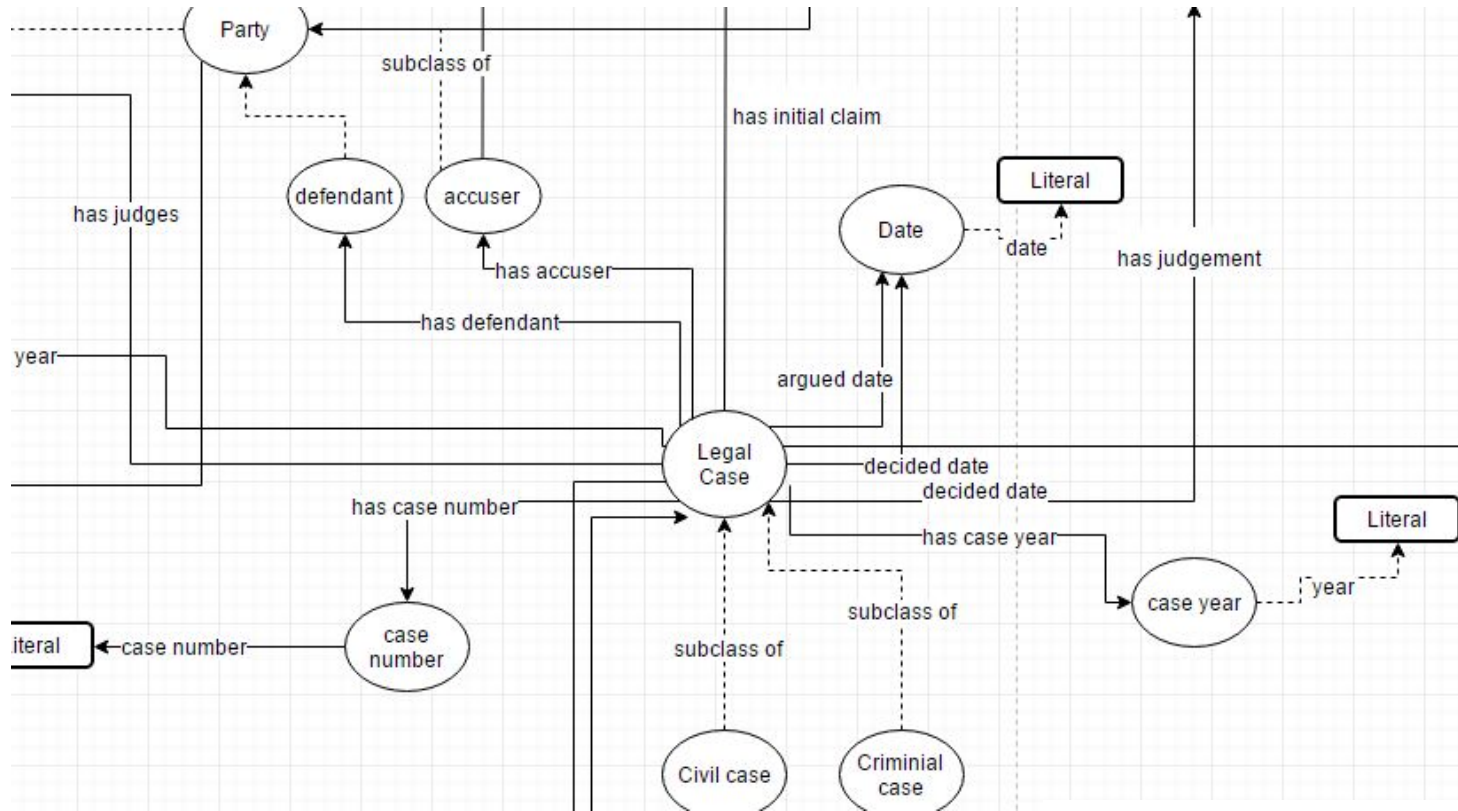
Ontology based Information Extraction

A system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies, to extract certain types of information and present the output using ontologies [4].

- Process unstructured or semi-structured natural language text
- Present the output using ontologies (generally)
- Use an information extraction process guided by an ontology

[4] D. C. Wimalasuriya and D. Dou , "Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches," *Journal of Information Science,* 2010.

Problem

1. The problem
2. What is missing ?

Problem

No proper effort has been taken for ontology driven legal information extraction based legal document retrieval.

Which results in,

Lack of a **easy to use**, **relevant,** and **legal domain specific** systems which enable **lawyers, paralegals,** and **general public** to get **relevant legal information** regarding a **court case of interest,** without having to use a **specific query language**.

Existing Systems ?

[5]

[6]
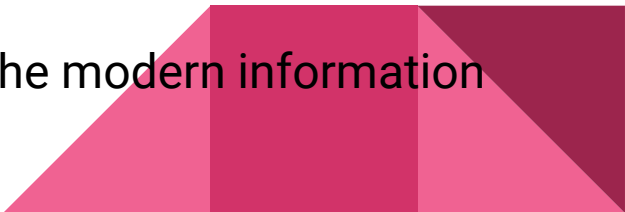
[7]

[5] T. Reuters, "FindLaw," Thomson Reuters, [Online]. Available: http://www.findlaw.com/. [Accessed 17 06 2017].
[6] T. Reuters, "WestLaw," Thomson Reuters, [Online]. Available: http://legalresearch.westlaw.co.uk/. [Accessed 17 06 2017].
[7] R. Group, "LexisNexis," RELX Group, [Online]. Available: https://www.lexisnexis.com/. [Accessed 17 06 2017].

What is Missing in Existing Systems ?

- **Ease of use**
  - Need to remember a query syntax (only experts can do it)

- **Can not handle natural human language as it is**
  - User should be aware of how to convert the information need, to a relevant query
  - General public is reluctant to use

- **More sophisticated underlying architecture to cater for more sophisticated information needs**
  - Need an underlying architecture which support the modern information needs

Motivation

1. Problems faced by legal professionals

2. Problems faced by general public

3. Problems faced by existing relevant systems

4. Ontology based Legal Information Extraction as a solution

Motivation

A Sample Legal Case [15]

John has developed a novel method for information extraction. The method is unique and it uses a machine learning algorithm. John tells his friend Peter about this project. John asks whether he can use Peter's data for testing. Peter tells that he will keep this confidential for friendship.

Peter does not sign a written agreement. Peter starts a startup with it.

John wants to sue Peter..

Problems Faced by Legal Professionals

- Having large number of legal documents

- Millions of legal cases

- New cases being added to the collection everyday

- Have to make very crucial decisions everyday

- A very busy occupation. Time is crucial

- Need to undergo extensive training to use helping systems such as WestLaw[6] and LexisNexis[7]

[6] T. Reuters, "WestLaw," Thomson Reuters, [Online]. Available: http://legalresearch.westlaw.co.uk/. [Accessed 17 06 2017].
[7] R. Group, "LexisNexis," RELX Group, [Online]. Available: https://www.lexisnexis.com/. [Accessed 17 06 2017].

Problems Faced by General Public

- Right to know the law [9]

- Inability to find legal resources without domain knowledge

[9] U. S. D. o. Justice, "FOIA.GOV," United States Department of Justice, [Online]. Available: https://www.foia.gov/. [Accessed 17 June 2017].

Problems Faced by Existing Systems

- User legal Information needs are changing rapidly

- Users seek more easy to use systems

- Systems should be exhibit feasibility for evolve

- Systems need to use modern day architectures to cater for needs

Ontology Based Legal Information Extraction

- Can address difficulties associated with legal domain specifically

- Can provide ease of use by providing ability to query using natural language

- Facilitate systems ability to evolve with time

- Can cope wp with users changing information needs

Objectives

1. Research
2. Research Objectives

Research

Use of Ontology based legal information extraction as a methodology for satisfying legal information need in the United States legal system, which will guide lawyers, paralegals and general public to get relevant legal information regarding a court case of interest, with both keyword and natural language based query processing.

Why USA

Research Objectives

- Gather and store relevant legal documents
  - Create a central data collection with all legal documents

- Create and store legal ontology
  - Customly make a ontology and store it

- Populate ontology with legal knowledge
  - Identification and implementation of ontology population method

- Defining efficient preprocessing techniques
  - Identify and incorporate relevant spell checkers, grammar checkers

- Information Extraction
  - Extract relevant information from natural text

Research Objectives

- Legal case retrieval query system
    - Create query generator
- Legal case retrieval
    - Retrieve legal cases for given information need

Technical Literature

1. Related work

Technical Literature

No	Work / Study	Relatedness	What's missing
1	D. A. d. ARAUJO, S. J. RIGO, C. Muller and R. Chishman, "**Information Extraction from Legal Documents Using Linguistic Knowledge and Ontologies**"	Usage of ontology to drive the legal information extraction process.	Only ten cases used. Manual term extraction process. Lack of preprocessing. Scalability issues due to due to heavy coupling with legal professionals to build the system.
2	S. B. Bruninghaus and K. D. Ashley, "**Improving the Representation of Legal Case Texts with Information Extraction Methods**," Seventh International Conference on Artificial intelligence and Law, 2001.	Addresses the need of improving the information representation from legal case texts with information extraction methods.	Based on text processing without using linguistic information and important relationship descriptions available in the domain knowledge of the legal context.
3	J. L. Kolodner - **An Introduction to Case-based Reasoning**	Case Based reasoning for legal cases	Only Text Processing Without linguistic information & relationship descriptions in the legal domain.

Technical Literature

No	Work / Study	Relatedness	What's missing
4	C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni and C. Soria, "**Automatic semantics extraction in law documents**"	'Multi Class Support Vector' - machine learning algorithm to do classification based information extraction	Performs poorly when compared to ontology based information extraction systems according to evaluation statistics.
5	P. Jackson, K. Al-Kofahi , C. Kreilick and B. Grom, "**Information extraction from case law and retrieval of prior cases by partial parsing and query generation**," CIKM '98 Proceedings of the seventh international conference on Information and knowledge management	Extracts rulings from court opinions and retrieves relevant prior cases from a citatory database, based on natural language processing techniques and statistical methods	Lack of domain knowledge incorporation
6	A. WYNER and W. PETERS, "**On Rule Extraction from Regulation**s," The 24th International Conference on Legal Knowledge and Information Systems (Jurix), 2011.	Extracts rulings by incorporating linguistic aspects in the design	Lack of domain knowledge incorporation. Based on NLP and Statistical methods only.

Technical Literature

No	Work / Study	Relatedness	What's missing
7	Walter S and Pinkal M. **Automatic extraction of definitions from german court decisions**	A rule based information extraction technique which extracts definitions from legal court decisions	Only for German Courts. Scalability issues. Limited number of definitions covered. Randomly selected verdicts.
8	J. R. Rene Witte, Ninus Khamis, "**Flexible ontology population from text: The owlexporter**."	Ontology Population using GATE	Heavy manual intervention. Not suitable for large ontologies
9	Andrew Carlson, Justin Betteridge, "**Toward an architecture for never ending language learning**"	Semi-supervised learning to never-ending language learning (NELL)	Only extract noun phrases. Defining rules for coupling. Results based on human interactions
10	R. S. Zhilin Yang, William W. Cohen, "**Revisiting semi supervised learning with graph embeddings**,"	A semi supervised learning methodology based on graph embeddings	Heavy weight onto predicts instances that are already observed in the graph in the training period, not for observe unlabeled data

Technical Literature

No	Work / Study	Relatedness	What's missing
11	Dimuthu Upeksha, Chamila Wijayarathna, Maduranga Siriwardena, Lahiru Lasandun, Chinthana Wimalasuriya, N.H.N.D. de Silva, and Gihan Dias, **Comparison Between Performance of Various Database Systems for Implementing a Language Corpus**	Cassandra as the ontology database.	-

Proposed Solution

1. Solution

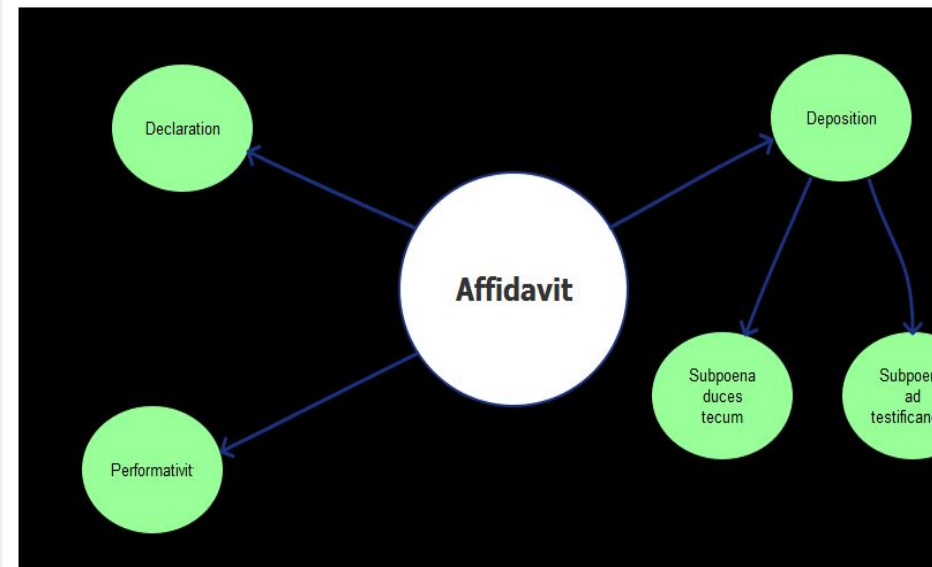# Key Concepts in the Solution

- Legal Ontology

- Information Extraction Module

- Pre-processor

- Ontology Population Module

- Legal Information Retrieval Module

- Query Generator Module

Proposed High Level Course of Actions

- Data (court cases, judgments and laws) collection using web crawlers from FindLaw [5] site.

- Create the legal ontology

- Improve and verify the ontology manually with domain experts

- Derive and develop the linguistic rules/ gazetteer lists from the ontology

- Extract information using the derived rules / lists

- Populate the ontology with legal knowledge

- Develop the retrieval of relevant cases based on the extracted information.

[5] T. Reuters, "FindLaw," Thomson Reuters, [Online]. Available: http://www.findlaw.com/. [Accessed 17 06 2017].

Enter item name

Declaration

Deposition

Affidavit

Subpoena duces tecum

Subpoena ad testificandum

Performativit

Definetion

An **affidavit** (/ˌæfɪˈdeɪvɪt/ *A-fə-DAY-vət*) is a written sworn statement of fact voluntarily made by an *affiant* or *deponent* under an oath or affirmation administered by a person authorized to do so by law.

Facts

Admission of an unsupported affidavit as evidence is unusual

Affidavits from persons who are dead or otherwise incapacitated, or who cannot be located or made to appear, may be accepted by t...

The acceptance of an affidavit by one society does not confirm its acceptance as a legal document in other jurisdictions.

[Expanded Definition] [Opinions] [Judgements]

Stephen Fells, owner and operator of a small convenience store in Milwaukee, Wisconsin, seeks review of a district court order upholding the final decision of the Food and Nutrition Service of the United States Department of Agriculture ("USDA"). The USDA decision permanently disqualified Mr. Fells's store, Stephen Snack Foods Candy & Variety, from participating in the Food Stamp Program, now referred to as the Supplemental Nutrition Assistance Program ("SNAP" or "Program"), for engaging in the trafficking of food stamps. Finding that Mr. Fells failed to satisfy his burden of proving by a preponderance of the evidence that the disqualification from SNAP was invalid, the district court upheld the agency's final determination of disqualification. On appeal, Mr. Fells continues to deny that he engaged in the trafficking of food stamps.

[Show mark up]

We are representing   [Stephen Fells ⌄]

[Generate Westlaw querey]

Trinity Homes LLC v. Ohio Cas. Ins. Co.   12/22/2010   09-3643
United States Seventh Circuit
Commercial Law, Construction, Contracts, Insurance Law
In a general contractors' suit against their insurers and an umbrella insurer seeking a declaration that all of the insurers had a duty to provide...

Research Automation, Inc. v. Schrader-Bridgeport Int'l, Inc. 11/23/2010 09-2232
United States Seventh Circuit
Civil Procedure, Commercial Law, Contracts, Remedies
In parties' mirror-image lawsuits filed in two different district courts involving a dispute between the seller and a buyer of an industrial cleaning...

Smart Mktg Group Inc. v. Publ'n Int'l Ltd.   10/28/2010   09-2646
United States Seventh Circuit
Commercial Law, Contracts, Cyberspace Law, Remedies
In a marketing firm's breach of contract suit against an online company, involving an agreement to develop programs that would deliver...

Reynolds v. CB Sports Bar, Inc.   10/22/2010   09-3753
United States Seventh Circuit
Commercial Law, Criminal Law & Procedure, Injury & Tort Law
In plaintiff's suit against a sports bar and two individual defendants for negligence and punitive damages, claiming that the two individuals induced...

US Supreme Court Briefs - FindLaw
http://supreme.findlaw.com/supreme_court/briefs/01-1120/01-1120-mer-ami-usa.html
... traffic in food stamp coupons, even where the store owner and manager have no knowledge of, and did not benefit from, the "trafficking" violation. ...

US Supreme Court Briefs - FindLaw
http://supreme.findlaw.com/supreme_court/briefs/00-1770/00-1770-pet-app.html
... from participation in the federal food stamp program ... or consent-illegally exchanged cash for food stamps. ... sharply in its favor." Big Country Foods, Inc ...

US Supreme Court Briefs - FindLaw
http://supreme.findlaw.com/supreme_court/briefs/04-1084/04-1084-mer-ja-v2.html
... for medical use by the Food and Drug ... using MAO inhibitors subsecuently ingested foods that contained ... an impression whereof is stamped in the ...

Comparson Of Prior Law And The Personal Responsibility And ...
http://corporate.findlaw.com/law-library/comparison-of-prior-law-and-the-personal-responsibility-and-work.html
... for second offense, permanently disqualifies individuals convicted of trafficking in food ... Food Stamps, continued, The Food Stamp Act contained ...

US Eighth Circuit: December 2010 Archives
http://blogs.findlaw.com/eighth_circuit/2010/12/
... with a witness to cover up the fraudulent use of food stamp benefits where ... of a firearm durnig and in relation to a drug trafficking offense, holding ...

Fells V. US, 10-1537 - US Seventh Circuit
http://blogs.findlaw.com/seventh_circuit/2010/12/fells-v-us-10-1537.html
... permanently disqualifing plaintiff's convenience store from participating in the Food Stamp Program for engaging in the trafficking of food stamps. ...

[Generate Case Documents]

Progress

1. Web Crawling
2. Ontology
3. Information Extraction
4. Ontology Storage
5. Implementations with Research Papers
   a. Ontology Population
   b. Legal information retrieval

Web Crawling and Data Collection

- Crawled cases, laws, judgments and executive orders [10] from,

  - FindLaw site

  - Justia site
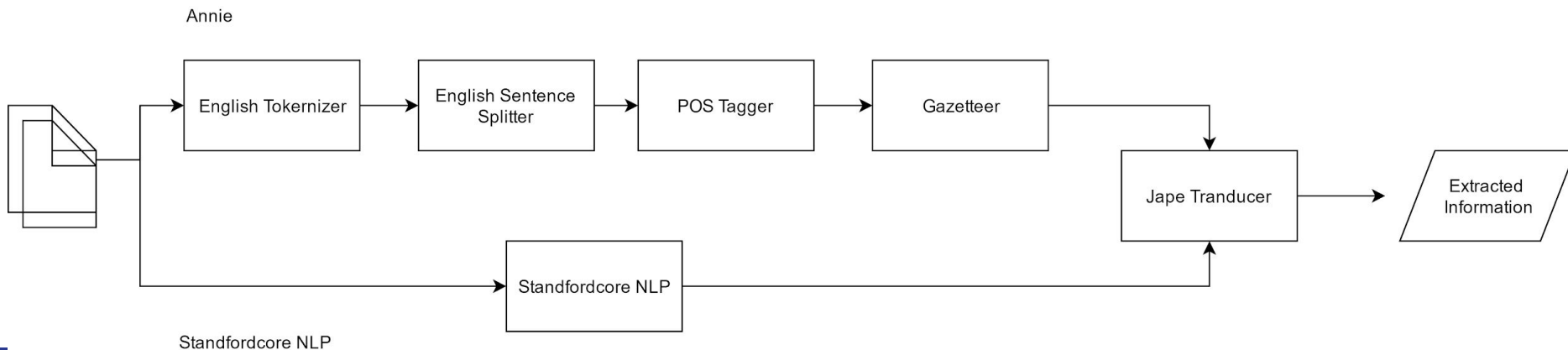
  - White House Executive Orders site

[10] G. 27, "Web Crawlers," Group 27, [Online]. Available: https://github.com/TeamTitanz/WebCrawler/tree/master/components. [Accessed 17 06 2017].

Ontology Creation and Verification

- Selected Legal Area : Consumer Protection Law

- Knowledge adding initial stage is completed

- Second stage of knowledge adding is still ongoing

- Manual verification from the domain experts

- [Demo]

Information Extraction

- JAPE is used for writing rules.

- Rules are being written based on the ontology.

- NLP Pipeline overview

  - Custom gazetteers are used

Annie

English Tokernizer → English Sentence Splitter → POS Tagger → Gazetteer → Jape Tranducer → Extracted Information

Standfordcore NLP

Standfordcore NLP

Research Papers

1. Deriving a Representative Vector for Ontology Classes with Instance Word Vector Embedding **(Published)**

2. Synergistic Union of Word2Vec and Lexicon for Domain Specific Semantic Similarity **(Accepted)**

3. Semi-Supervised Instance Population of an Ontology using Word Vector Embeddings **(Accepted)**

4. Similarity Based Document Embeddings for Legal Information Retrieval **(Submission : September 1st)**

5. A Survey of the Known Approaches of Information Extraction in the Legal Domain (**Work in progres**s)

6. Legal Document Retrieval Accuracy Measuring System (**Upcoming**)

Deriving a Representative Vector for Ontology Classes with Instance Word Vector Embedding

- State : **Published**
- Conference : **Seventh International Conference on Innovative Computing Technology (INTECH 2017)***
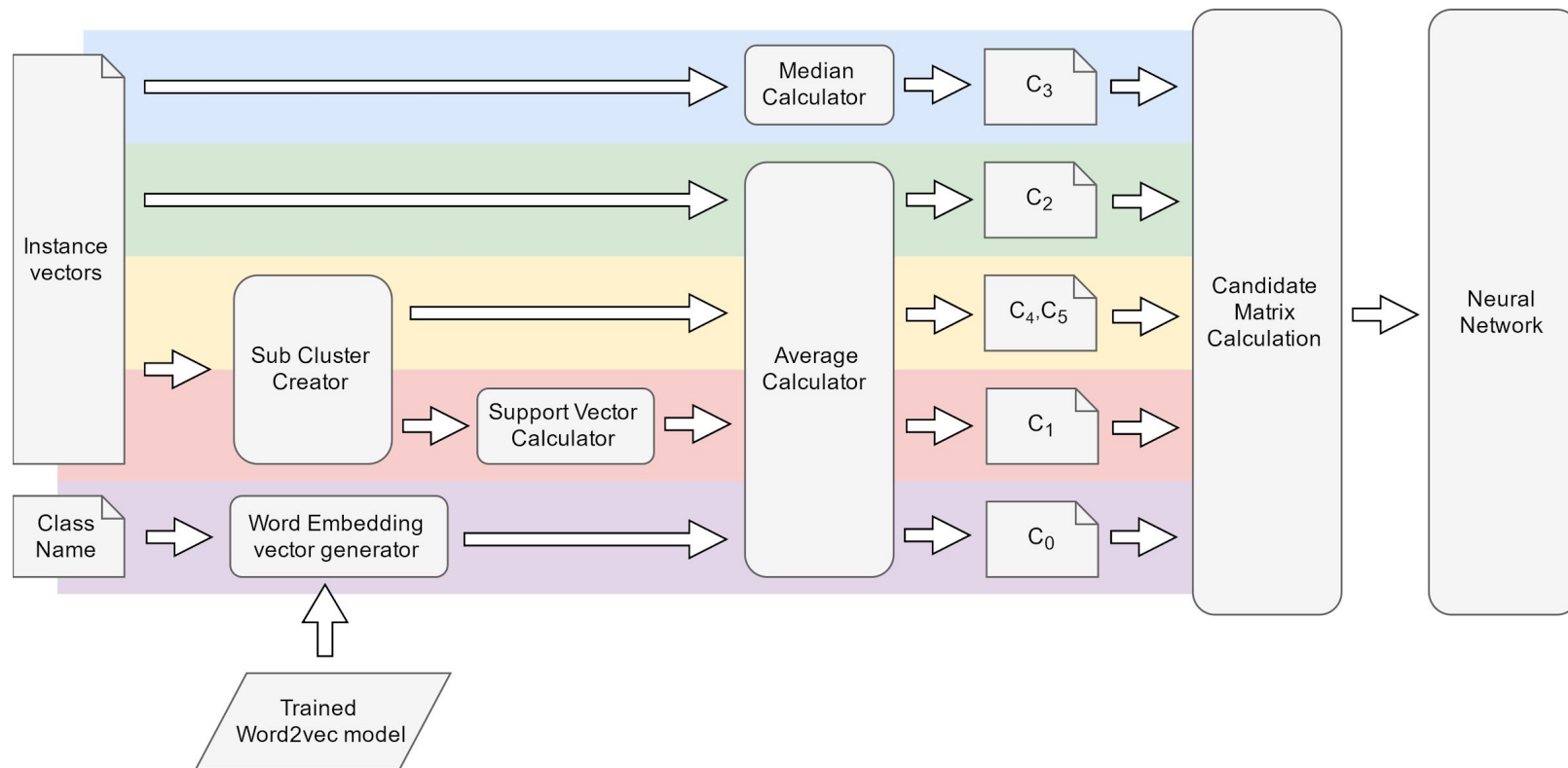- August 16-18, 2017 - Luton, UK

* http://www.dirf.org/intech/

Methodology for deriving instance vectors
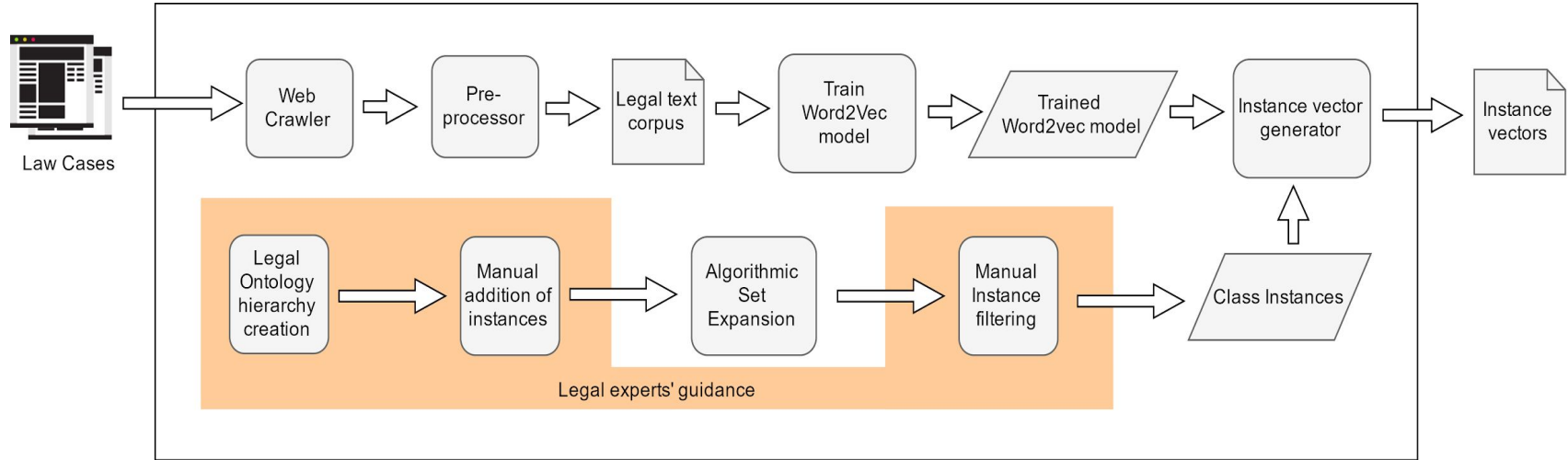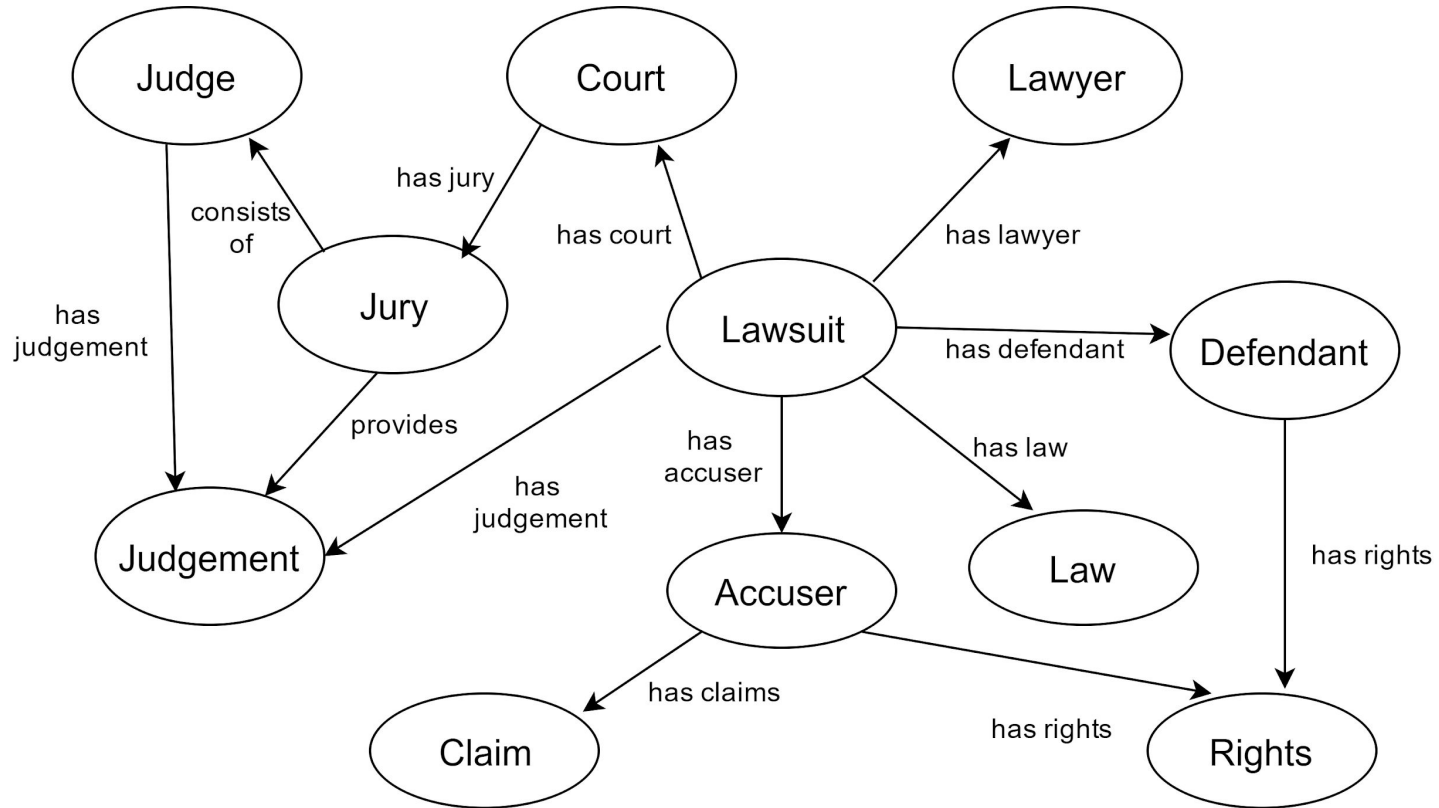
Methodology for Training and Testing

Semi-Supervised Instance Population of an Ontology using Word Vector Embeddings

- State : **Accepted**
- Conference : **International Conference on Advances in ICT for Emerging Regions ***
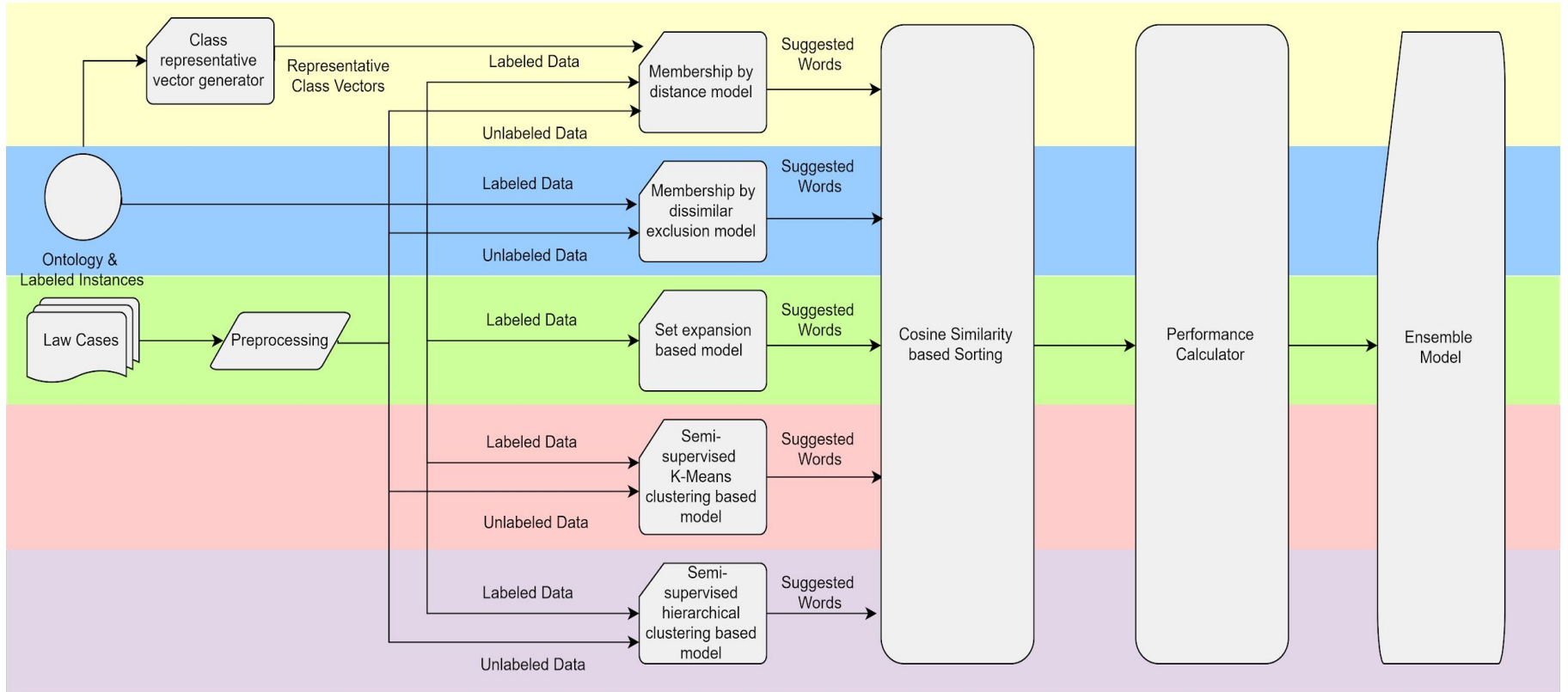- 7th - 8th September, 2017

* http://www.icter.org/conference/

# Sub Ontology Used

Flow Diagram of the Methodology

A Survey of the Known Approaches of Information Extraction in the Legal Domain

- State : **Work in Progress (75% completed)**

Synergistic Union of Word2Vec and Lexicon for Domain Specific Semantic Similarity

- State : **Accepted**
- Conference : **12th IEEE International Conference on Industrial & Information Systems 2017 ***
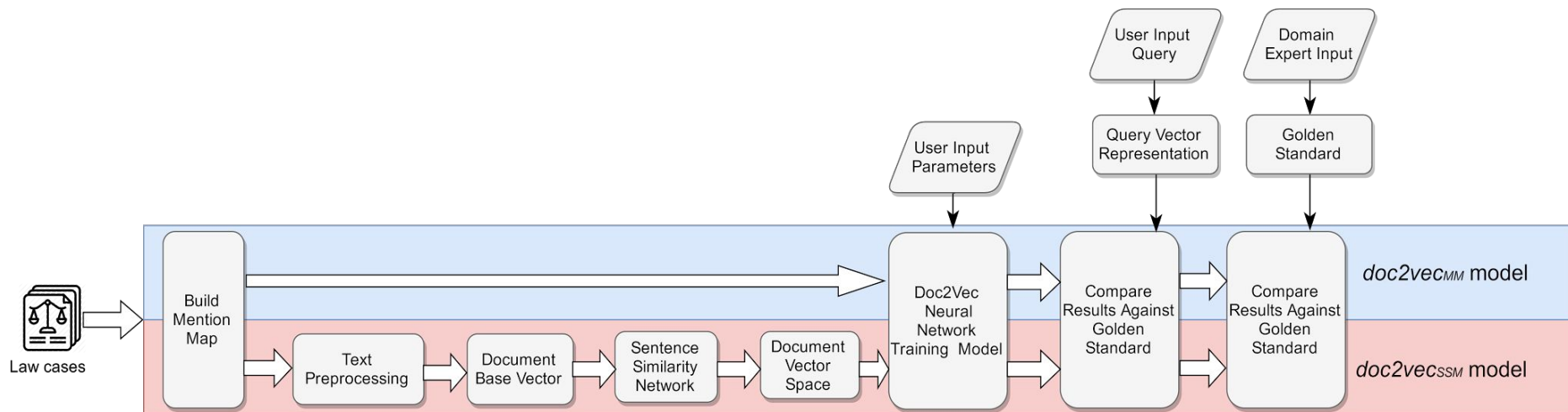- Dec 15, 2017 - Dec 16, 2017

Overall Methodology
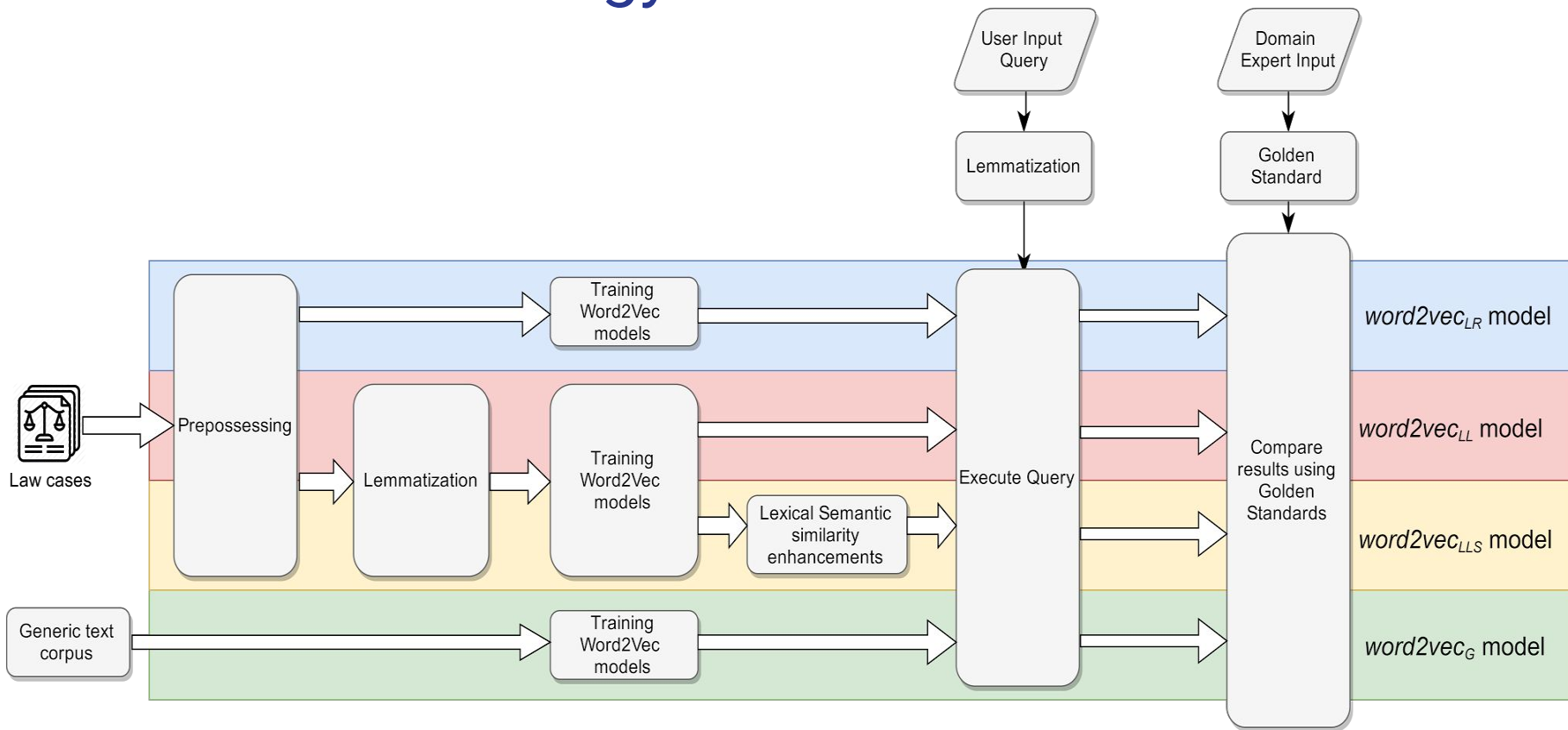
Similarity Based Document Embeddings for Legal Information Retrieval

- State : **Await Submission**

Methodology

Thank you !!

Students - Group 27

V.M.Jayawardana
K.M.Sugathadasa
B.A.Rathnayaka
K.J.T.D.Lakmal

Supervisors

Dr. Shehan Perera
Mr. Nisansa de Sliva

References

[1]. S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, pages 848-850, 2003.

[2] T. Gruber. Ontolingua: A Translation Approach to Providing Portable Ontology Specifications. Knowledge Acquisition, 5(2):199–220, 1993.

[3] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. Data Knowledge Engineering, 25(1-2):161–197, 1998.

[4] D. C. Wimalasuriya and D. Dou , "Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches," *Journal of Information Science,* 2010.

[5] T. Reuters, "FindLaw," Thomson Reuters, [Online]. Available: http://www.findlaw.com/. [Accessed 17 06 2017].

[6] T. Reuters, "WestLaw," Thomson Reuters, [Online]. Available: http://legalresearch.westlaw.co.uk/. [Accessed 17 06 2017].

[7] R. Group, "LexisNexis," RELX Group, [Online]. Available: https://www.lexisnexis.com/. [Accessed 17 06 2017].

References

[9] U. S. D. o. Justice, "FOIA.GOV," United States Department of Justice, [Online]. Available: https://www.foia.gov/. [Accessed 17 June 2017].

[10] G. 27, "Web Crawlers," Group 27, [Online]. Available: https://github.com/TeamTitanz/WebCrawler/tree/master/components. [Accessed 17 06 2017].

Extra Slides

1. Q & A

Why United States Legal System ?

All the Legal Documents are available online

Countries like Sri Lanka has hard copies (need additional effort in the project)

Fully focus on the computing aspects

Can extend to any Legal system available online

back

WestLaw Query Language

Information need: Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company. **Query**: "trade secret" /s disclos! /s prevent /s employe!

Information need: Requirements for disabled people to be able to access a workplace. **Query**: disab! /p access! /s work-site work-place (employment /3 place)

Information need: Cases about a host's responsibility for drunk guests. **Query**: host! /p (responsib! liab!) /p (intoxicat! drunk!) /p guest
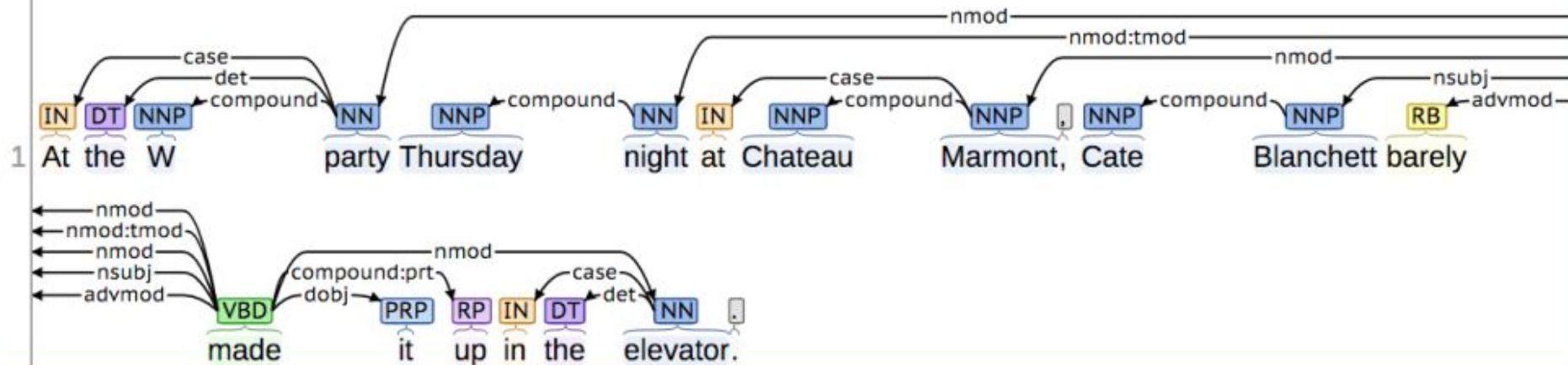
NLP Pipeline.
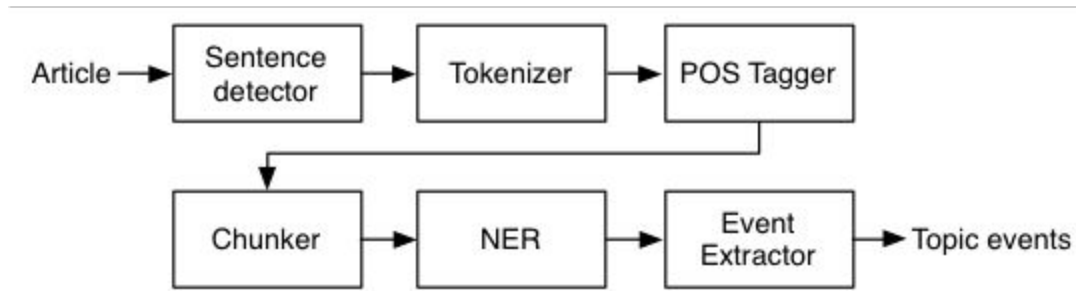
Annotations

Named Entity Recognition:

1 At the W party [Date Thursday] [Time night] at [Location Chateau Marmont], [Person Cate Blanchett] barely made it up in the elevator.

Basic Dependencies:

Gate Annotation Structure

	Text			
	Cyndi savoredthe soup.			
	^0...^5...^10..^15..^20			
	Annotations			
Id	Type	SpanStart	Span End	Features
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	

POS Tagging

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
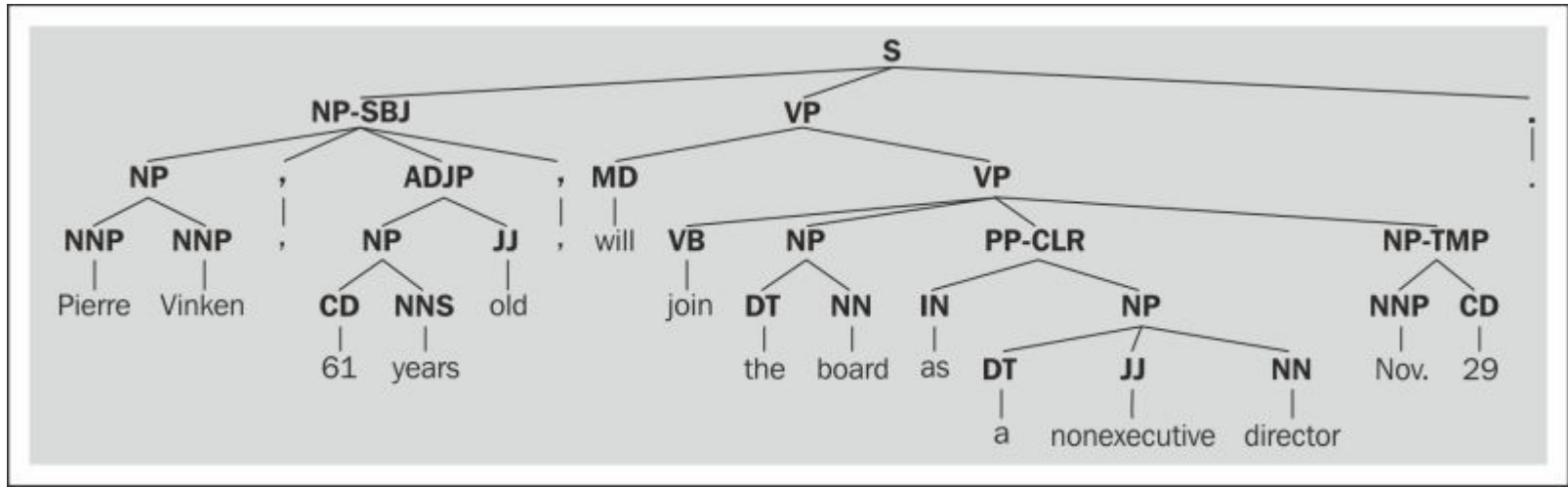
OBIE General Architecture [4]

[4] D. C. Wimalasuriya and D. Dou , "Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches," *Journal of Information Science,* 2010.