Hybrid Information Extraction Systems for Open Data

Class 1: Introduction

Pablo Ariel Duboue, PhD

Curso de Posgrado Cs. de la Computacion
FaMAF-UNC

What is Information Extraction?

- Information Extraction (IE): Natural Language Processing (NLP) right now
- A series of techniques to extract unstructured information from text into structured data.
  - Very limited semantic content

Full Semantic Analysis

"The young boy quickly threw a green rubber ball at a big fast car."

```
(make-frame clause1
  (clauseid (value clause1))
  (propositionid (value proposition1))
  (speechactid (value speech-act1)))

(make-frame proposition1
  (propositionid (value proposition1))
  (clauseid (value clause1))
  (process-type (value action))
  (is-token-of (value *throw))
  (aspectid (value aspect1))
  (space (value proposition1.space))
  (time (value (at proposition1.time)))
  (subworld (value everyday-world))
  (modality (value real))
  (manner (value quickly))
  (agent (value role1))
  (object (value role2))
  (instrument (value role3))
  (source (value role1))
  (destination (value role4)))

(make-frame aspect1
  (phase (value end))
  (iteration (value 1) )  ;in seconds
  (duration (value 1)))
```

```
(make-frame speech-act1
  (speech-act (value assertion))
  (direct? (value no))
  (speaker (value author))
  (hearer (value reader))
  (time (value speech-act1_time)))
(make-frame role1
  (roleid (value role1))
  (clauseid (value clause1))
  (comment (value "a young boy"))
  (referent (value person1))
  (description (value role1))
  (is-token-of (value *person))
  (age (value (10 11)))
  (sex (value male)))

(make-frame role2
  (roleid (value role2))
  (clauseid (value clause1))
  (comment (value "a green rubber ball"))
  (referent (value ball1))
  (description (value role2))
  (is-token-of (value *ball))
  (color (value green))
  (made-of (value *rubber)))
```

```
(make-frame role3
  (roleid (value role3))
  (clauseid (value clause1))
  (comment (value *arm*))
  (referent (value arm1))
  (description (value role3))
  (is-token-of (value *arm))
  (part-of (value role1)))

(make-frame role4
  (roleid (value role4))
  (clauseid (value clause1))
  (comment (value "a big fast car"))
  (referent (value vehicle1))
  (description (value role4))
  (is-token-of (value *vehicle))
  (medium (value road))
  (propulsion (value fuel))
  (wheels (value 3  4 ) )
  (size (value big))
  (velocity (value (50 180) ) ) )
```

From "Seleccion Lexica en Traduccion Automatica" (Duboue, to appear) adapted from Nirenburg and Nirenburg (1998)

Motivating example

- Imran, Muhammad, et al. "Practical extraction of disaster-relevant information from social media." Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013.
  - 4-page paper describing a very circumscribed IE system over Twitter using only ML
- Tornado hits town of Joplin, 2011 (206.7k tweets)
- Hurricane Sandy 2012 (140k tweets)
- Find informative tweets and extract information depending on the class:
  1. Caution and Advice
  2. Casualties and Damage
  3. Donations of money, goods or services
  4. People missing, found, or seen
  5. Information Sources
  6. Other

Irman et al. (2013) example

- People class, missing or lost people

  *rt @911buff: public help needed: **2 boys 2 & 4 missing** nearly 24 hours after they got separated from their mom when car submerged in si. #sandy #911buff*

- Casualties and damage: Infrastructure

  *RT @TIME: NYC building had numerous construction complaints before **crane collapse** http://t.co/7EDmKOp3 #Sandy*

- Casualties and damage: Injured or dead

  *At least **39** dead millions without power in Sandy's aftermath. http://t.co/Wdvz8KK8*

Irman et al. (2013) example

- Donations: Requests money/goods/services

  400 Volunteers are needed *for areas that #Sandy destroyed.*

- Donations: Offers money/goods/services

  I want to volunteer to ***help the hurricane Sandy victims****. If anyone knows how I can get involved please let me know!*

Imran et al. (2013) techniques

- Two steps:
  - Cascade of Naive Bayes classifiers: personal / informative (direct, indirect) / unrelated-to-disaster
    - Uses Naive Bayes classifier from Weka. 2000 / 4400 annotated tweets using annotators.
    - We will discuss classifiers tomorrow (class 2)
  - CRFs for IE (class 4)
    - Uses sequence labeling implemented with Conditional Random Fields (CRFs)
    - We will discuss them on Thursday (class 4)
- Annotation using an on-line provider
  - Simple instructions
  - Cut&Paste interface
  - We will discuss annotations on Friday (class 5)

Imran et al. (2013) metrics

- Imran et al. (2013) example:

	Annotated	System Output
(a)	There were **12** injured	<NIL>
(b)	A **bridge has collapsed**	bridge
(c)	**10 volunteers** needed	needed

- Detection rate ("there's data in the Tweet": 66% {b,c} over {a,b,c})
  - binary document recall over slots
  - very unusual but useful for their task
- Hit ratio ("some of the extracted data is good": 50% {b} over {b,c})
  - precision using overlap matching
  - usual metric, although overlap match is tolerated by few tasks

Imran et al. (2013) results

- Joplin: 78% / 90%
- Sandy: 41% / 79%
- But, train on all Joplin then test on Sandy: **11%** / 78% *(does not generalize)*
- Train on Joplin + 10% Sandy, test on remainder Sandy: 21% / 81% *(transfer learning)*
- How to further improve?
  - Incorporate domain heuristics using rules
  - We will discuss IE rules tomorrow and Wednesday (classes 2 and 3)

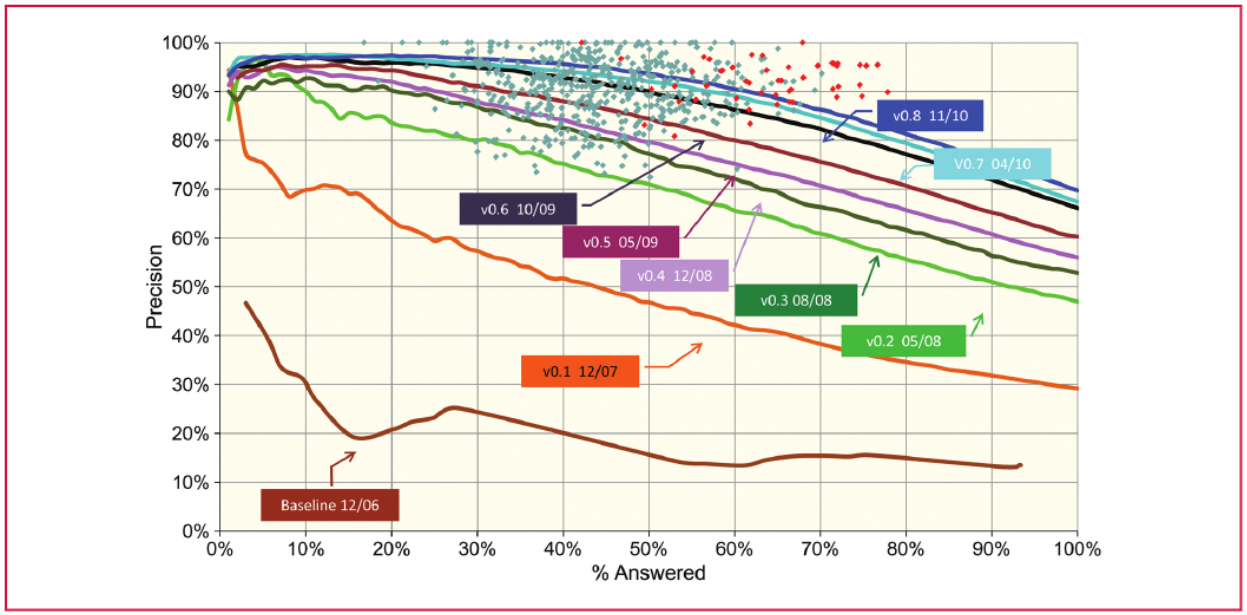
Hybrid IE

- You will encounter plenty of research on doing IE using only ML
- This makes sense on a research level
  - Solving a problem by programming does not get a paper published
  - Unless you are publishing to the community that had that problem (i.e., biology if doing IE on biology texts)
- However:
  - Quality annotations are also expensive
  - Combining rules and machine learning is more an art than a science
  - It is hard to build intuitions on the topic

The First Step Goes The Longer

- IE as many other tasks in NLP is fast to get to a certain performance
  - Further improvement requires an amount of effort multiple times the original
    - And for much less absolute performance
  - On Friday (class 5) we will discuss an adaptation task that took:
    - 3hs to produced a system 79%
    - +9hs (total 12hs) to take it to 80%

The First Step Goes The Longer (cont.)



From Ferrucci et al. (2012).

Applications of IE

- Traditional applications:
  - Question Answering (QA)
  - Information Retrieval (IR)
    - CiteSeer
  - Report Generation

Applications of IE (cont.)

- *Incorporating unstructured sources into statistical models*
- Automatic inferencing
- From Text Mining, Weiss et al. (2005) Chapter 6: Information Extraction
  - Criminal justice (drug networks)
  - Intelligence (Non-Obvious Relationship Awareness)
- Automated decision systems

Not all errors are the same depending on the application

Specificity and Complexity Performance Tradeoff

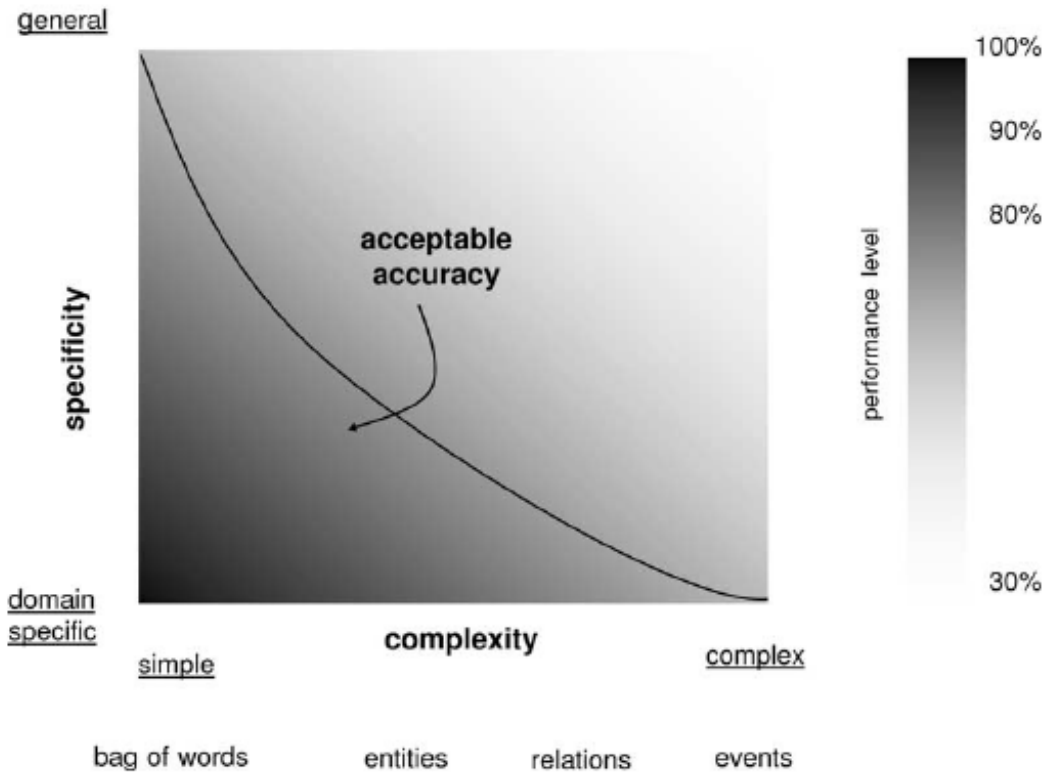


Figure 1 of Cunningham, H (2006)

Metrics

- Measuring how many times a system outputs the right answer ("accuracy") is not enough
  - Many interesting problems are very biased towards a background class
  - If 95% of the time something doesn't happen, saying it'll never happen (not a very useful classifier!) will make you only 5% wrong
- Metrics:

$$precision = \frac{|correctly\ tagged|}{|tagged|} = \frac{tp}{tp + fp}$$

$$recall = \frac{|correctly\ tagged|}{|should\ be\ tagged|} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

Precision / Recall Trade-off

- If you never answer anything, you are being very precise
  - 100% precision, not very useful
  - But what about returning one output out of 10, but all correct?

Building a KB with correct information

- If you return every subsequence of characters, you are returning all valid answers
  - But what about returning 9 out of 10, without missing any correct?

Pre-filtering the input

Matching Outputs to Human Annotations

- Another issue with evaluation, what to do with partial outputs?
  - *The contract was awarded to [Acme] Inc.*
  - vs
  - *The contract was awarded to [Acme Inc.]*
  - What about *"The contract was awarded to [International Business] Machines"*?
- And outputs with spurious material?
  - *The contract was awarded [to Acme Inc.]*
  - But then *"The contract was award to [Acme Associates a limitied partnership from Ottawa]"*?
- Again, this is application dependent

About Pablo

- Universidad Nacional de Cordoba - FaMAF
  - Trabajo Final de Licenciatura: "Desarrollo de un Parser Funcional para el Lenguaje Castellano", 1998
  - ECI 1995 - Dr. Juan Garay "Adversarios y Computacion"
- Columbia University
  - Doctoral Dissertation: "Indirect Supervised Learning of Strategic Generation Logic", defended Jan. 2005.
- IBM Research Watson
  - Deep QA - Watson - Jeopardy! Show
- In Montreal from 2010-2016
  - Two semesters teaching at Cordoba University (NLG / ML on large datasets)
- On a personal sabbatical in NY now

IE and me

- Started my PhD back in 1999 working on GeneWays
  - A multidisciplinary IE pipeline for genomics
- Did my PhD in Natural Language Generation
  - Intelligence domain (IE from news)
- At IBM: Enterprise Search competition (2006)
  - Expert search (expert detection and linking)
- After IBM worked on two IE projects
  - Real Estate contracts in French
  - Technical support from Web pages

GeneWays: biology

- First year of PhD
- Same term used to refer to gene / protein /mRNA
- So ambiguous human authors disambiguate it themselves in some cases
  - Use those cases to train a classifier
- Conference paper (my most cited one, 234 citations)
  - "Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach"
    Hatzivassiloglou, Duboue and Rzhetsky
    Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology. 2001
- Journal (my most cited one, 270 citations)
  - GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data
    Andrey Rzhetsky et al. (12 authors)
    Journal of Biomedical Informatics Vol. 37 (1) 2004

ProGenIE: intelligence

- Profile Generation by Information Extraction
- Wrote general purpose extractors using GATE and JAPE
- Assemble information about a person from news articles, then build a biography
- Focus was on the generation bit, IE aspect was proof of concept
- Financed by the AQuAinT program
- Conference paper
  - "ProGenIE: Biographical descriptions for Intelligence Analysis" Duboue, McKeown, and Hatzivassiloglou. Proceedings of the NSF/NIJ Symposium on Intelligence and Security Informatics. Lecture Notes in Computer Science 2003

IBM: Expert Search

- TREC Enterprise Search competition (2006)
- IBM Team
- Collect contexts around the names of experts to gaugue their expertise on a topic
- Use the same contexts to disambiguate occurrences (entity detection and linking)

Real Estate

- At a Montreal consulting firm specialized in IE
- Real estate contracts OCR'ed
- High performance, high availability pipeline

Technical Support

- At a Montreal company specialized in technology for Tech Support centers
- Identify pages with tech support content
- Extract devices, versions, problems and solutions

IE Competitions

- Best way to gauge which ideas work and which ideas doesn't is to agree upon a task (including a dataset) and compare results
- Necessary when not all errors contribute the same and the systems are built based on heuristics
- Pioneered by IE since the late 80s

  MUC 80s-90s, navy, single document

  ACE 2000s, multilingual, focus on the entities rather than documents

  KBP multidocument, focus on assembling a high quality knowledge base

MUC

- Follow the discussion in "Evaluating Message Understanding Systems: An Analysis of MUC-3" by Chincor et al. in Comp Ling. (1993)
- MUC-3: 15 systems over terrorism in Latin America

  train 1300 documents (400k word tokens over 18k word types

  test 3 sets of 100 documents each (dry run, official test, muc-4)
- Three approaches:
  1. Pattern-matching
  2. Syntax-driven
  3. Sematic-driven

MUC-3 Example

TST2-MUC3-0069

BOGOTA, 7 SEP 89 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [MARIBEL OSORIO] [TEXT] MEDELLIN CONTINUES TO LIVE THROUGH A WAVE OF TERROR. FOLLOWING LAST NIGHT'S ATTACK ON A BANK, WHICH CAUSED A LOT OF DAMAGE, A LOAD OF DYNAMITE WAS HURLED AGAINST A POLICE STATION. FORTUNATELY NO ONE WAS HURT. HOWEVER, AT APPROXIMATELY 1700 TODAY A BOMB EXPLODED INSIDE A FAST-FOOD RESTAURANT.

A MEDIUM-SIZED BOMB EXPLODED SHORTLY BEFORE 1700 AT THE PRESTO INSTALLATIONS LOCATED ON [WORDS INDISTINCT] AND PLAYA AVENUE. APPROXIMATELY 35 PEOPLE WERE INSIDE THE RESTAURANT AT THE TIME. A WORKER NOTICED A SUSPICIOUS PACKAGE UNDER A TABLE WHERE MINUTES BEFORE TWO MEN HAD BEEN SEATED. AFTER AN INITIAL MINOR EXPLOSION, THE PACKAGE EXPLODED. THE 35 PEOPLE HAD ALREADY BEEN EVACUATED FROM THE BUILDING, AND ONLY 1 POLICEMAN WAS SLIGHTLY INJURED; HE WAS THROWN TO THE GROUND BY THE SHOCK WAVE. THE AREA WAS IMMEDIATELY CORDONED OFF BY THE AUTHORITIES WHILE THE OTHER BUSINESSES CLOSED THEIR DOORS. IT IS NOT KNOWN HOW MUCH DAMAGE WAS CAUSED; HOWEVER, MOST OF THE DAMAGE WAS OCCURRED INSIDE THE RESTAURANT. THE MEN WHO LEFT THE BOMB FLED AND THERE ARE NO CLUES AS TO THEIR WHEREABOUTS.

Figure 4 of Chincor et al. (1993)

- Documents come from a DB query to the FBIS

(Argentina OR Bolivia OR Chile OR Colombia OR Ecuador OR (El Salvador) OR Guatemala OR Honduras OR Peru) AND (abduct OR abduction OR ambush OR arson OR assassinate OR assassination OR assault OR (blow [up]) OR bomb OR bombing OR explode OR explosion OR hijack OR hijacking OR kidnap OR kidnapping OR kill OR killing OR murder OR rob OR shoot OR shooting OR steal OR terrorist)

MUC-3 Example (cont.)

- Translated from Spanish by foreign broadcast information service (FBIS)
  - Text is all uppercase, strange and idiosyncratic
  - Text is interspersed with quoted speech and guerrilla communiques
  - Bad spelling and mistranslations
- 1,600 documents
  - 1,300 development set
  - 300 annotated:
    - 100 dev
    - 100 test MUC-3
    - 100 test MUC-4

MUC-3 Example Template

0. MESSAGE ID	TST2-MUC3-0069
1. TEMPLATE ID	1
2. DATE OF INCIDENT	(06 SEP 89) / (06 SEP 89 - 07 SEP 89)
3. TYPE OF INCIDENT	ATTACK
4. CATEGORY OF INCIDENT	? TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	-
6. PERPETRATOR: ID OF ORG(S)	-
7. PERPETRATOR: CONFIDENCE	-
8. PHYSICAL TARGET: ID(S)	"BANK"
9. PHYSICAL TARGET: TOTAL NUM	1
10. PHYSICAL TARGET: TYPE(S)	FINANCIAL: "BANK"
11. HUMAN TARGET: ID(S)	-
12. HUMAN TARGET: TOTAL NUM	-
13. HUMAN TARGET: TYPE(S)	-
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPES(S)	-
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17. EFFECT ON PHYSICAL TARGET(S)	SOME DAMAGE: "BANK"
18. EFFECT ON HUMAN TARGET(S)	-

- Template has 18 slots
  - Some slots reference other slots
- Different template types have different slots required
  - Murder has no physical-target, those are for attack

MUC-3 Example Template (cont.)

0. MESSAGE ID	TST2-MUC3-0069
1. TEMPLATE ID	2
2. DATE OF INCIDENT	07 SEP 89
3. TYPE OF INCIDENT	BOMBING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"TWO MEN" / "MEN"
6. PERPETRATOR: ID OF ORG(S)	-
7. PERPETRATOR: CONFIDENCE	-
8. PHYSICAL TARGET: ID(S)	"FAST-FOOD RESTAURANT" / "PRESTO INSTALLATIONS" / "RESTAURANT"
9. PHYSICAL TARGET: TOTAL NUM	1
10. PHYSICAL TARGET: TYPE(S)	COMMERCIAL: "FAST-FOOD RESTAURANT" / "PRESTO INSTALLATIONS" / "RESTAURANT"
11. HUMAN TARGET: ID(S)	"PEOPLE" "POLICEMAN"
12. HUMAN TARGET: TOTAL NUM	36
13. HUMAN TARGET: TYPE(S)	CIVILIAN: "PEOPLE" LAW ENFORCEMENT: "POLICEMAN"
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPES(S)	*
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17. EFFECT ON PHYSICAL TARGET(S)	SOME DAMAGE: "FAST-FOOD RESTAURANT" / "PRESTO INSTALLATIONS" / "RESTAURANT"
18. EFFECT ON HUMAN TARGET(S)	INJURY: "POLICEMAN" NO INJURY: "PEOPLE"

Figure 5 of Chincor et al. (1993)

MUC-3 Challenges

- Problems:
  - Deal with noisy, real world data (unusual for 1992)
  - Discriminate by date, old vs new information, criminal vs terrorist
    - Not all documents are relevant
- Annotating the output templates was challenging
- Deciding what to do with spurious templates (tiered evaluation):
  1. matched only (considerate only 1 slot, template-id for the wrong templates)
  2. matched/missing (for spurious, consider only template-id but take all for missing, official metric)
  3. all templates (all missing slots and all spurious slots, official metric for muc-4)

MUC-3 Generic System Architecture

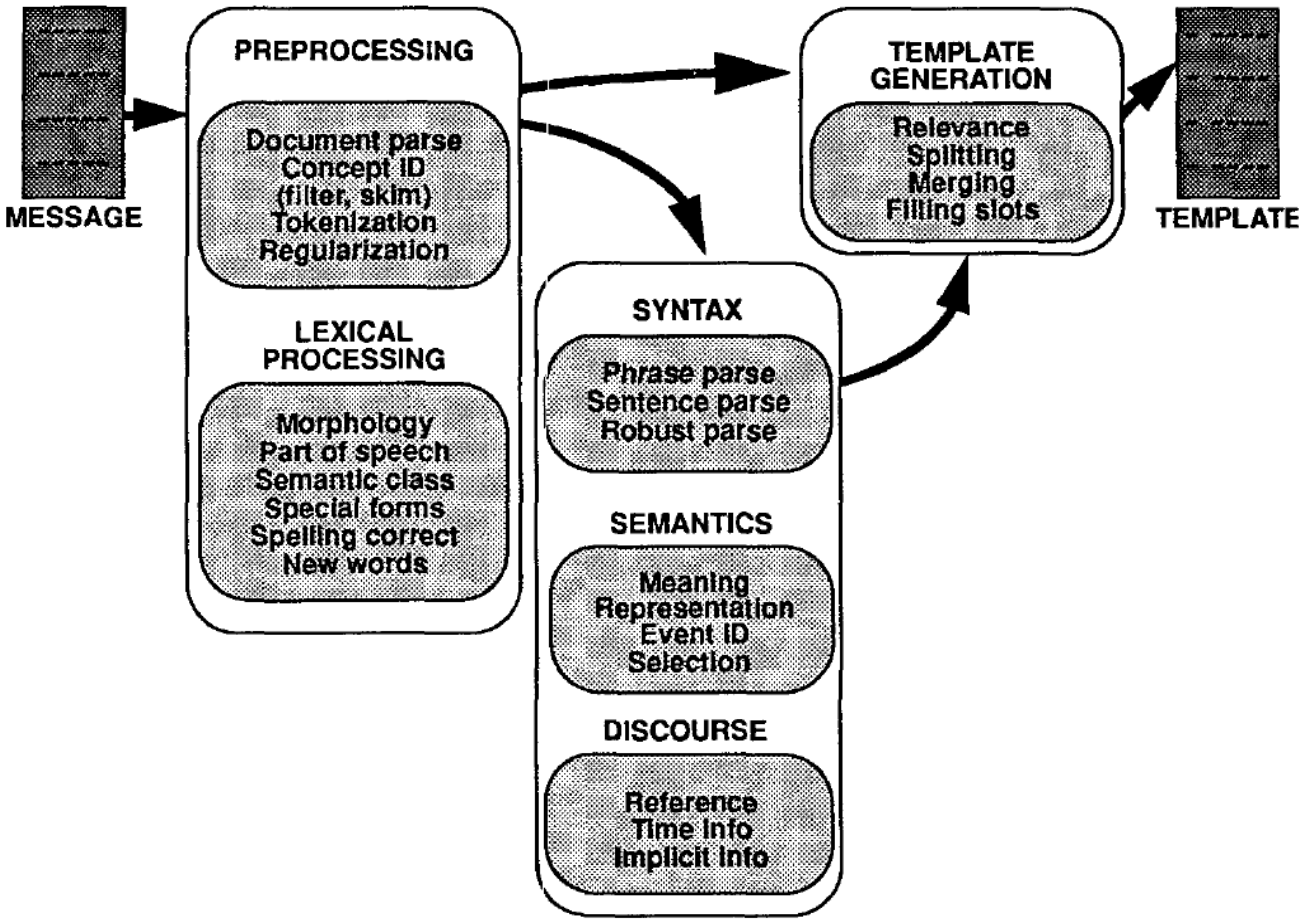


Figure 9 of Chincor et al. (1993)

Precision / Recall Trade-off at MUC-3

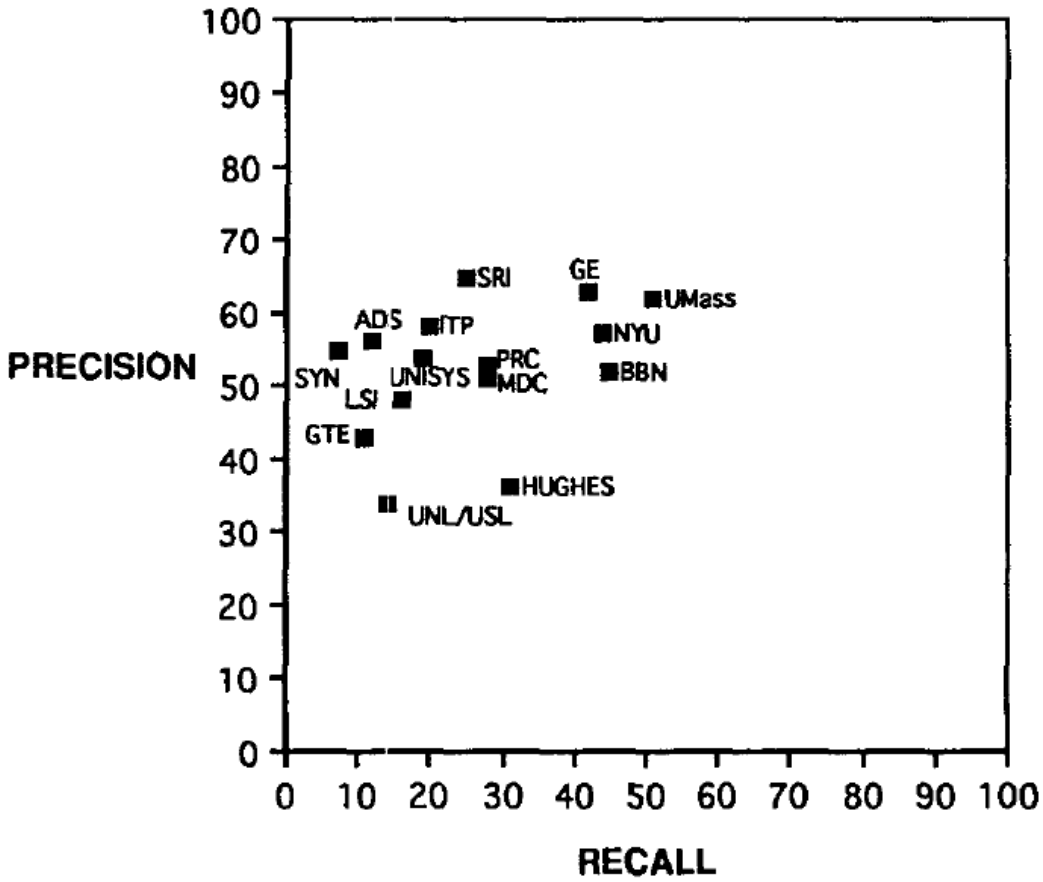


Figure 2 of Chincor et al. (1993)

MUC-3 Conclusions

- Sites reported 10-11 person months for domain migration/adaptation.
- ML team, 3 month effort
- Hybrid: 6 month effort and top performer

Other MUCs

Conference	Year	Text Source	Topic (Domain)
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

from Wikipedia

ACE

- This discussion follows "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation" by Doddington et. al. (2004) at LREC
- ACE objective:
  - Infer the entities being mentioned, relations among those entities directly expressed and events where the entities participate
  - Multilingual (English, Arabic, Chinese) / multimodal (speech, scanned documents)
  - Extracting information from transduced (ASR/OCR) output text

2000-2001 entities

2002-2003 entities + relations

2004-2008 entities + relations + events

ACE Approach

- Data, lots of data. Enough data to attempt ML approaches.

Corpus/ Project Phase	Data Amount (words/language)	Tasks	Languages
ACE-Pilot	15K training	entities	English
ACE-1	180K training, 45K evaluation	entities	English
ACE-2	180K training, 45K dev, 45K eval	entities, relations	English, Chinese
ACE 2003	100K training, 50K evaluation	entities, relations	English, Chinese, Arabic
ACE 2004	300K training, 50K evaluation	entities, relations, events	English, Chinese, Arabic

Table 1 of Doddington et. al. (2004)

ACE Subtasks

EDT Entity Detection and Tracking

- Main task
- All entity mentions (names, descriptions and pronouns) need to be clustered
- 7 entity types (Person, Organization, Location, Facility, Weapon, Vehicle and Geo-Political Entity) (+ subtypes)
- Nested mentions are captured ([Facility [PER George Washington] Bridge])

RDC Relation Detection and Characterization

- 5 types (Role,Part,At,Near,Social) / 24 subtypes of relations
- Explicit relations vs. contextual relations (no world knowledge relations)

ACE Subtasks (cont.)

VDC Event Detection and Characterization

- Generalization of relations
- Type (destroy, create, transfer, move, interact) and modality (real, not real)

LNK Entity Linking

- Coreference

ACE Challenges

- Complex task for annotators
  - Require annotators that know linguistics
  - Difficulty of keeping teams in different languages in sync
  - 5%-10% of the data re-annotated by different people
- Some interannotator agreement is really low (English relations are at 35%)

TAC

- This discussion follows "Overview of the TAC 2009 Knowledge Base Population Track" by McNamee et al. (2009) at TAC.
- KBP: learn things from a corpus.
  - ACE: document based, KBP: corpus based
  - Learning wrong things is **bad**
- Datasets: 1.3M English newswire + blogs collection from LDC (6.5Gb), from 2007-2008
  - 200k people
  - 200k GPEs
  - 80k org
  - 300k misc from Wikipedia '08

TAC Subtasks

- Two tasks:
  1. Entity Linking
     - Linking target entities to their Wikipedia page
     - Annotated 560 entities, 3904 mentions
  2. Slot filling (learning attributes about targets)
     - 53 entities.
- Top performing system, 82% accuracy

TAC Hardest Queries

- Subsidiary organization
  - Xinhua Finance Ltd .vs Xinhua Finance Media Ltd
- Typographical mistake / ambiguous acronym
  - DCR for Democratic Republic of Congo
  - MND (Taiwan Ministry of National Defense) referred to as NDM in text
- Metaphorical ´names´
  - Iron Lady (several strong female politicians)
- Unclear referent
  - New Caledonia (country or soccer team)
- Mistakes in assessments
  - NYC Dept of Health, not US Dept of Health
  - NY State Dept of of Health, not US Dept of Health

from McNamee et al. (2009), page 30

IE Subtasks

- This discussion follows "Information Extraction: Capabilities and Challenges" by Ralph Grishman at the International Winter School in Language and Speech Technologies (2012)

1. Entity detection / linking
2. Event detection
3. Frame building

Named Entity Recognition

- NER as a task was introduced in MUC-6
  - Useful outside of IE, eg MT, QA
- Now 200 categories and no longer "names" (e.g., colors)
  - Initially high performance REs for names
- Example:

[Fred Flintstone]$^{\text{person}}$ was named [CTO]$^{\text{position}}$ of [Time Bank Inc.]$^{\text{organization}}$ in [2031]$^{\text{date}}$. The [next year] [he] got married and became [CEO]$^{\text{position}}$ of [Dinosaur Savings & Loan]$^{\text{organization}}$.

Named Entity Recognition Techniques

- We will discuss them in class 2 (tomorrow)
  - Regular Expressions
  - Lists of names (gazzetteers)
  - Machine learning

Linking Names to Entities

- In-document coreference
  - Pronouns detection is good (80-90%)

  [Fred Flintstone] was named CTO of Time Bank Inc. in 2031. The next year [he] ...
  - Nouns coreference is not that good

  Fred Flintstone was named CTO of Time Bank Inc. in [2031]. The next [year] he ...
- Cross-document coreference in research
- Techniques:
  - Name matching
  - Word Sense Disambiguation
- We will discuss them in class 2 (tomorrow)

Relations

- Distinguish events (n-ary) vs relations (binary)
- Example relations from TAC 2004:
  - Cross-sentence needed in 15% of cases

relation type	subtype
physical	located, near, part-whole
personal-social	business, family, other
employment / membership / subsidiary	employ-executive, employ-staff, employ-undetermined, member-of-group, partner, subsidiary, other
agent-artifact	user-or-owner, inventor-or-manufacturer, other
person-org	affiliation ethnic, ideology, other
GPE affiliation	citizen-or-resident, based-in, other
discourse	-

Relation Detection Techniques

- We will discuss them in classes 3 and 4
  - Hand-crafted patterns (generalising using word classes and dependency trees)
  - Classifier: sentence + entities $\Rightarrow$ relation (yes/no)

Events and Scenarios

- Relations with more than two slots
- ACE 2005 events:

Event type	Subtypes
Life	Be-born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-ownership, Transfer-money
Business	Start-org, Merge-org, Declare-bankruptcy, End-org
Conflict	Attack, Demonstrate
Personnel	Start-position, End-position, Nominate, Elect
Justice	Arrest-jail, Release-parole, Trial-hearing Charge-indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

Events and Scenarios Techniques

- Single topic documents ⇒ scenario templates
  - Hand crafted (MUC):
    1. Slot filling
    2. Consolidation
  - Supervised:
    - Trigger-based classification (we will discuss on class 5 when we discuss recent research)
- IRE: Implicit Relation Extraction, when the document describes a unique event
  - Focus of this course, we find a subdocument that contain an IRE then use sequence tagging to extract it

This Course

- We will focus on high quality extraction from highly paradigmatic texts
  - Very, very laborious
  - Plenty of industrial applications / interest
  - We will see a full example in a framework (other frameworks will be mentioned)
- We will not directly cover some topics directly
  - No general news / Twitter extraction
    - Many of the techniques apply but the error rate is much higher
  - No Open IE (relations signaled by lexical items)
    - Interesting idea, but require extra (human) work for many traditional applications
- Discuss recent research directions on Friday (class 5)

Target Students

- Students curious about NLP research in general
- NLP students interested in moving into IE research
- Open Data afficionados
- Professionals with a need for industrial IE
- Analytics professionals

Current Research

- On class 5 (Friday) we will discuss current research in the following topics:
  - Multilinguality: extracting information from texts written in multiple languages (ACE)
  - Multiple sources: extracting information from multiple texts involving the same entities (KBP)
  - Character-based Deep Learning models
  - Applying IE to novel media (i.e., social)
  - Modeling multiple domains at the same time (Open Domain IE)

NLP Pipelines

- As an IE system can profit from most of the results of NLP analytics, a high performance IE system usually involves performing a number of processing stages over the text
- These stages build on previous stages and are usually assembled as a pipeline of components

Linguistics Primer

- POS
- Dependencies
- Parsing: subject / objects / verbs
- Phrases
- Semantic classes

Part-of-Speech

- Nouns: entities in the discourse
- Articles: transform nouns into noun phrases
- Adjectives: transform nouns into nouns
- Adverbs: transform verbs into verbs
- Verbs: transform noun phrases (and others) into verb phrases
- Noun phrases: transform verb phrases into sentences
- Pronouns: substitute for a noun or a noun phrase

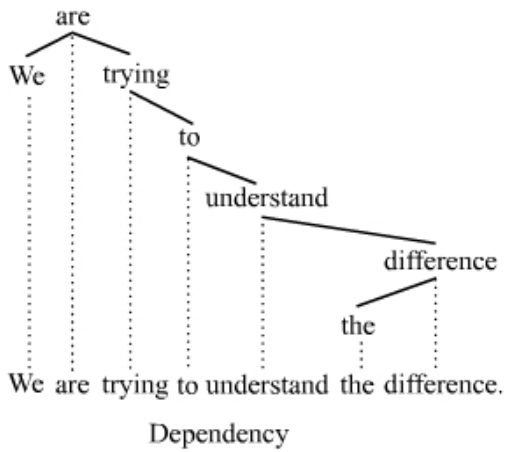
See https://en.wikipedia.org/wiki/Categorial_grammar

Part-of-Speech Tagging

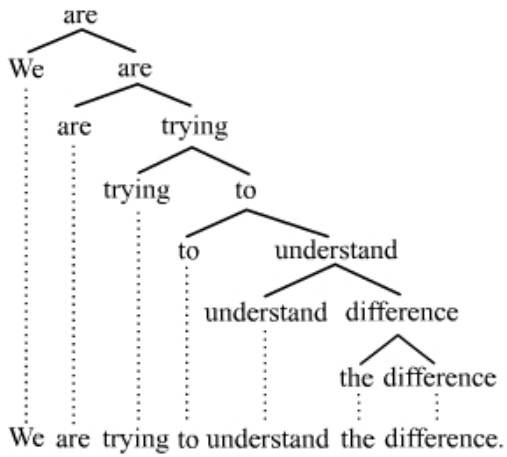
- Assign a morphosynactic class to each word in a sentence
- It is challenging because many words have multiple potential uses as different classes
  - Then I saw the **fire**.
  - She thinks they might **fire** him.
- A "solved" task in NLP
  - Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
  - "97.24%" accuracy per token
  - 56% **sentence** level accuracy
- That's including a class like /TO

Dependency Parsing

- Instead of building a full syntactic parse, associate each word for a "head" word



Dependency



Constituency (BPS)

from Wikipedia

Full Parsing

- Identify the different constituents of the sentence
  - Noun phrases (their core concept, the head, is a noun)
  - Verb phrases: verb head
  - Adjectival phrases
  - Adverbial phrases
- Different theories of parsing, different grammars
- Most systems use a Tree Bank of human annotated parse trees
  - Wall Street Journal articles

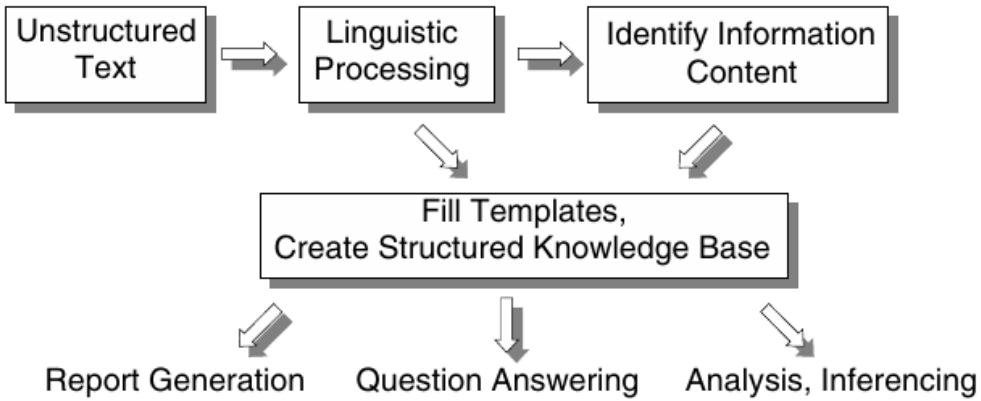
Role Labeling

- A type of semantic parsing
- For certain verbs, identify "roles" (list adapted from Natural Language Annotation for Machine Learning by Pustejovsky and Stubbs (2013))

agent	participant that is doing or causing the action to occur
theme/figure	participant who undergoes a change in position or state
experiencer	participant experiencing or perceiving something
source	location or place from which motion begins; person from whom the theme is given
goal	location or place to which the motion is directed or terminates

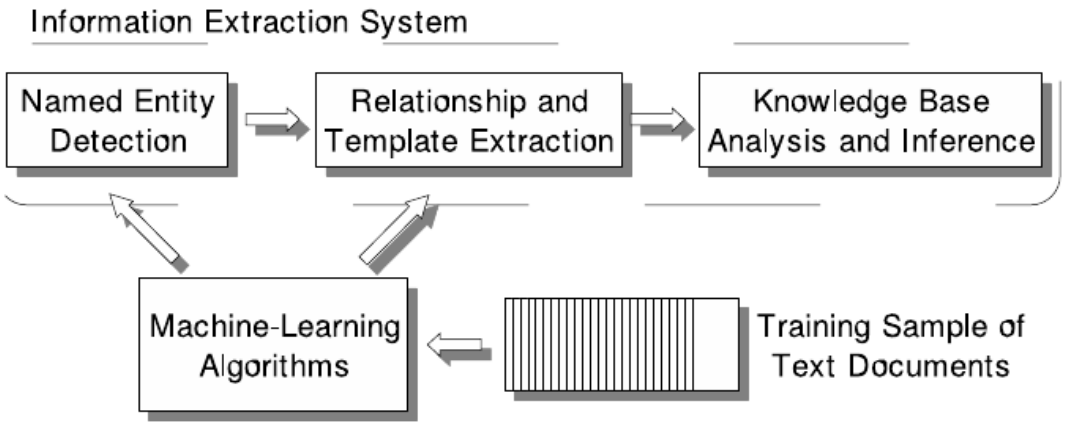
(...)

IE Pipelines



From Text Mining, Weiss et al. (2005) Figure 6.2

IE Pipelines (cont.)



From Text Mining, Weiss et al. (2005) Figure 6.3

What is UIMA

- UIMA is a framework, a means to integrate text or other unstructured information analytics.
- Reference implementations available for Java, C++ and others.
- An Open Source project under the umbrella of the Apache Foundation.

Analytics Frameworks

- Find all telephone numbers in running text
  - `(((\([0-9]{3}\))|[0-9]{3})-?[0-9]{3}-?[0-9]{4}`
- Nice but...
  - How are you going to feed this further processing?
  - What about finding non-standard proper names in text?
  - Acquiring technology from external vendors, free software projects, etc?

In-line Annotations

- Modify text to include annotations
  - This/DET happy/ADJ puppy/N
- It gets very messy very quickly
  - (S (NP (This/DET happy/ADJ puppy/N) (VP eats/V (NP (the/DET bone/N)))
- Annotations can easily cross boundaries of other annotations
  - He said <**confidential**>the project can't go own. The funding is lacking.</**confidential**>

Stand-off Annotations

- Standoff annotations
  - Do not modify the text
  - Keep the annotations as offsets within the original text
- Most analytics frameworks support standoff annotations.
- UIMA is built with standoff annotations at its core.
- Example:

```
He said the project can't go own.  The funding is lacking.
---------------------------------------------------------
012345678901234567890123567890123456789012345678901234567890123456789012345678
```

  - Sentence Annotation: 0-33, 36-58.
  - Confidential Annotation: 8-58.

Type Systems

- Key to integrating analytic packages developed by independent vendors.
- Clear metadata about
  - Expected Inputs
    - Tokens, sentences, proper names, etc
  - Produced Outputs
    - Parse trees, opinions, etc
- The framework creates an unified typesystem for a given set of annotators being run.

Many Frameworks

- Besides UIMA
  - `http://uima.apache.org`
- LingPipe
  - `http://alias-i.com/lingpipe/`
- Gate
  - `http://gate.ac.uk/`

UIMA Advantages

- Apache Licensed
- Enterprise-ready code quality
- Demonstrated scalability
- Developed by experts in building frameworks
  - Not domain (e.g., NLP) experts
- Interoperable (C++, Java, others)

Java Refresher

- Java is a statically typed compiled to virtual machine code language
  - Class-based, single inheritance
- Each class becomes a dynamically linked library loaded at runtime by the virtual machine
- Also the base of the Dalvik VM used in Android
- Dependency management and build cycle is usually managed with the Apache Maven tool:
  - mvn clean
  - mvn compile
  - mvn test
  - mvn deploy
  - pom.xml (Project Object Model)

UIMA tutorial

- Common Annotation Structure or CAS
  - Subject of Analysis (SofA or View)
  - JCas
- Feature Structures
  - Annotations
- Indices and Iterators
- Analysis Engines (AEs)
  - AEs descriptors

Room Annotator

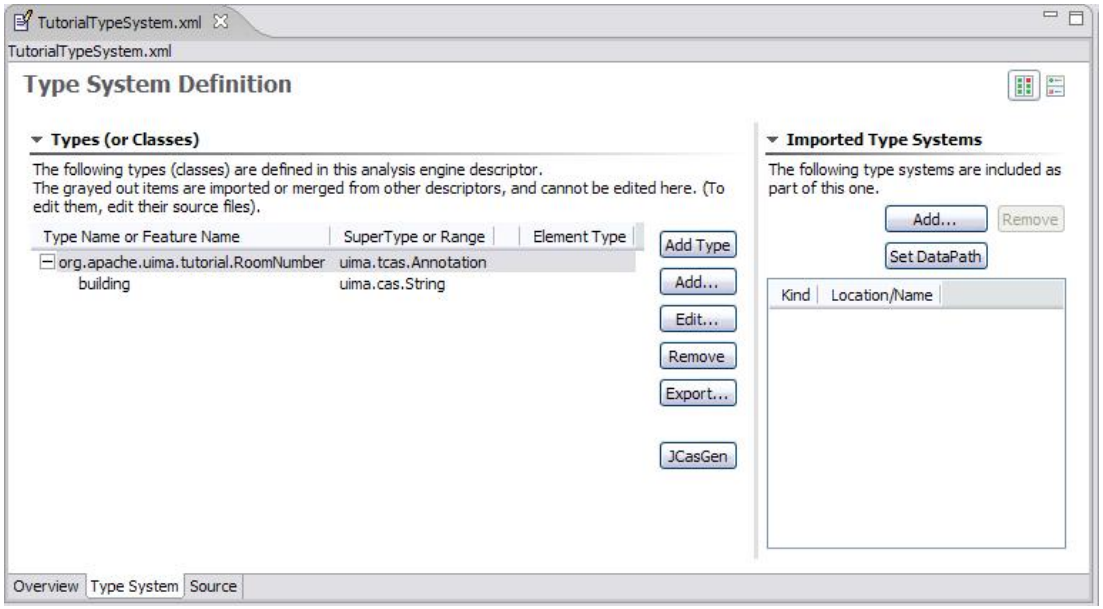
- From the UIMA tutorial, write an Analysis Engine that identifies room numbers in text.

  Yorktown patterns: 20-001, 31-206, 04-123 (Regular Expression Pattern: [0-9][0-9]-[0-2][0-9][0-9])

  Hawthorne patterns: GN-K35, 1S-L07, 4N-B21 (Regular Expression Pattern: [G1-4][NS]-[A-Z][0-9])

- Steps:
  1. Define the CAS types that the annotator will use.
  2. Generate the Java classes for these types.
  3. Write the actual annotator Java code.
  4. Create the Analysis Engine descriptor.
  5. Test the annotator.

Editing a Type System



From UIMA Tutorial and User's Guide

The XML descriptor

```
<?xml version="1.0" encoding="UTF-8" ?>
  <typeSystemDescription xmlns="http://uima.apache.org/resourceSpecifier">
    <name>TutorialTypeSystem</name>
    <description>Type System Definition for the tutorial examples -
        as of Exercise 1</description>
    <vendor>Apache Software Foundation</vendor>
    <version>1.0</version>
    <types>
      <typeDescription>
        <name>org.apache.uima.tutorial.RoomNumber</name>
        <description></description>
        <supertypeName>uima.tcas.Annotation</supertypeName>
        <features>
          <featureDescription>
            <name>building</name>
            <description>Building containing this room</description>
            <rangeTypeName>uima.cas.String</rangeTypeName>
          </featureDescription>
        </features>
      </typeDescription>
    </types>
  </typeSystemDescription>
```

The AE code

```
package org.apache.uima.tutorial.ex1;

import java.util.regex.Matcher;
import java.util.regex.Pattern;

import org.apache.uima.analysis_component.JCasAnnotator_ImplBase;
import org.apache.uima.jcas.JCas;
import org.apache.uima.tutorial.RoomNumber;

/**
 * Example annotator that detects room numbers using
 * Java 1.4 regular expressions.
 */
public class RoomNumberAnnotator extends JCasAnnotator_ImplBase {
  private Pattern mYorktownPattern =
        Pattern.compile("\\b[0-4]\\d-[0-2]\\d\\d\\b");

  private Pattern mHawthornePattern =
        Pattern.compile("\\b[G1-4][NS]-[A-Z]\\d\\d\\b");

  public void process(JCas aJCas) {
    // next slide
  }
}
```

The AE code (cont.)

```
public void process(JCas aJCas) {
  // get document text
  String docText = aJCas.getDocumentText();
  // search for Yorktown room numbers
  Matcher matcher = mYorktownPattern.matcher(docText);
  int pos = 0;
  while (matcher.find(pos)) {
    // found one - create annotation
    RoomNumber annotation = new RoomNumber(aJCas);
    annotation.setBegin(matcher.start());
    annotation.setEnd(matcher.end());
    annotation.setBuilding("Yorktown");
    annotation.addToIndexes();
    pos = matcher.end();
  }
  // search for Hawthorne room numbers
  // ..
}
```

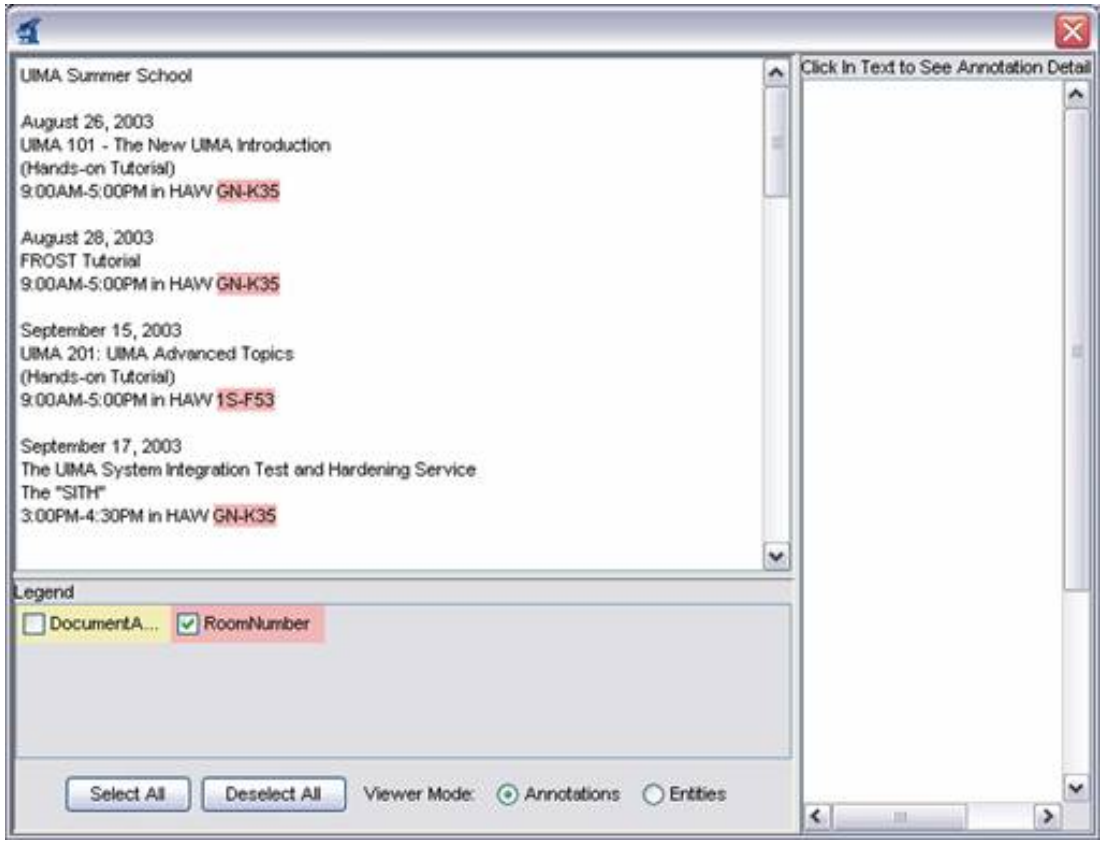
An UIMA Maven Project

```
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/
xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xs
<modelVersion>4.0.0</modelVersion>

...

                <dependency>
                        <groupId>org.apache.uima</groupId>
                        <artifactId>uimaj-core</artifactId>
                        <version>2.7.0</version>
                </dependency>
...
```

UIMA Document Analyzer

From UIMA Tutorial and User's Guide

UIMA Document Analyzer (cont.)



From UIMA Tutorial and User's Guide

Control Flow

- UIMA allows you to specify which AE will receive the CAS next, based on all the annotations on the CAS.
- `examples/descriptors/flow_controller/WhiteboardFlowController.xml`
  - FlowController implementing a simple version of the "whiteboard" flow model. Each time a CAS is received, it looks at the pool of available AEs that have not yet run on that CAS, and picks one whose input requirements are satisfied. Limitations: only looks at types, not features. Does not handle multiple Sofas or CasMultipliers.

Annotation Serialization

- XMI is an OMG standard for expressing objects graphs in XML.

```
<?xml version="1.0"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi=http://www.omg.org/XMI
xmlns:cas="http:///uima/cas.ecore" xmlns:myproj="http:///org/myproj.ecore">
<cas:Sofa xmi:id="1" sofaNum="1"
text="the quick brown fox jumps over the lazy dog."/>
<myproj:Foo xmi:id="2" begin="14" end="19" myFeature="bar"/>
<cas:View sofa="1" members="2"/>
</xmi:XMI>
```

End-to-End Case Study

- Octroy Pipeline
  - Information Extraction for Open Data
  - Proceedings of the Executive Committee in the cities of Montreal and Laval
  - French
  - Focus on finding which company got what money for what reason

Open Data and IE

- Open Data: governments realising large quantities of unstructured information
- Use IE to fulfill Open Data mission of increased awareness and transparency

The Data

- FIOA request by Montreal Gazette journalist Roberto Rocha
  - 28,376 decisions
  - 112Mb
  - 458,190 lines
  - 2,456,420 words
  - 87 words per decision (average)
  - 25k pages if printed
- NEQ: Enterprise Registry of Quebec (CC-BY-NC)
  - 2,088,934 companies

The Problem

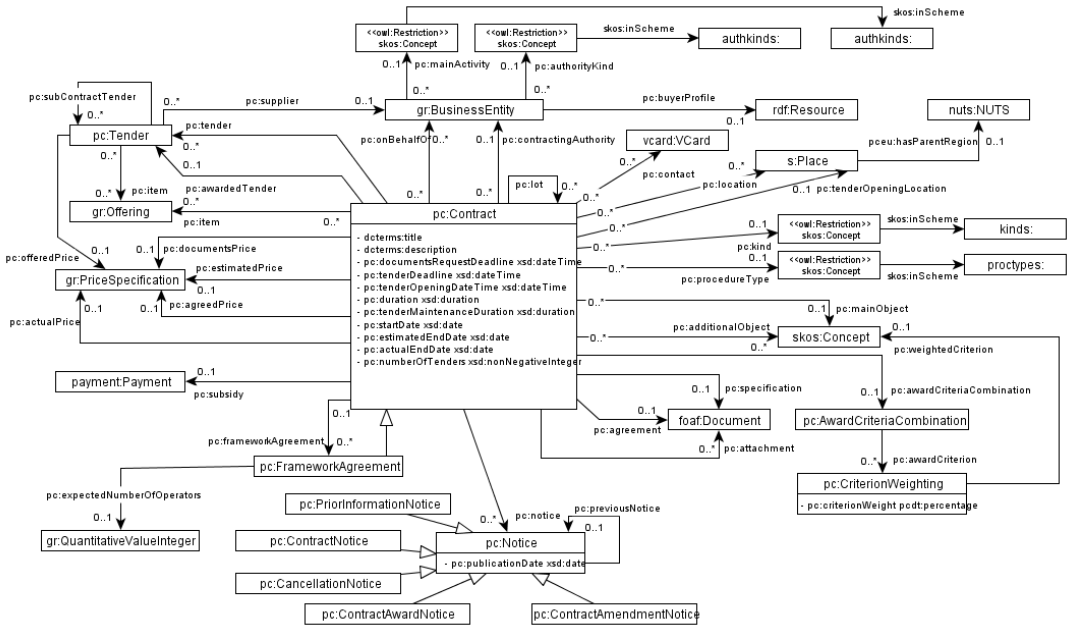
- Identify which decisions authorize (French *octroyer*) payment to companies
- For these decisions, extract company name, amount and reason
- Link company to the NEQ registry

Types

- Types
  1. Enterprise
  2. Amount
  3. Reason
- Relation
  1. Contract

Output

- https://github.com/opendatacz/public-contracts-ontology



Octroy Components

1. Collection reader (runs from the DB and brings up the date, too)
2. Relevant text segments (RuTA)
3. Sentence splitting (OpenNLP)
4. Tokenization (OpenNLP)
5. Amount annotation (RuTA)
6. Enterprise dictionary (using NEQ dictionary)
7. CRF (for Enterprise and Reason)
8. Postprocessing (RuTA)
9. Entity linking (to the NEQ ID)
10. Event extraction (populates the .cz ontology)
11. CAS Consumer (RDF writer)

Initial Data Analysis

- Data Release Schedule:
  - randomize document IDs, pick 30% as development set
  - 8,513 as devset
- Take the first 36 and analyze them by hand
  - 18 contains contracts (50%)
- Issues:
  - Approving money transfer to another gov't org (not a contract)
  - Contracts with no amount (kept)
  - Subsidies (kept as contract, the reason is subsidy)
  - When a company *pays* the gov't it has all the information as when the *gov't* pays a company.

Examples

Contrat de construction
CM Montréal-Nord , Bureau du directeur
d'arrondissement - 1063602020
Octroyer un contrat de 1 328 000,65 $, taxes incluses à TGA Montréal inc. pour des travaux de réfection d'égout, d'aqueduc, de trottoirs et de pavages sur les avenues Drapeau, Éthier et Patricia (Contrat No 755)

Examples (cont.)

CE-2011/4460
CONTRAT - SOUMISSION «OS-ING/2011-041» ÉQUATION GROUPE CONSEIL INC.
RÉSOLU À L'UNANIMITÉ: que la soumission «OS-ING/2011-041» déposée par la firme Équation groupe conseil inc. concernant les services professionnels d'ingénieurs-conseils pour la préparation des plans et devis ainsi que pour les services durant la construction des travaux d'aménagement d'un terrain de tennis double au parc des Nénuphars, prévus au règlement L-11796, soit acceptée et qu'à cette fin, la firme susdite prépare les plans, devis et documents de soumissions au montant de 31 317,07 $; que les honoraires soient calculés conformément aux dispositions de la soumission «OS-ING/2011-041»; que le Greffier ou la Greffière adjointe retourne les garanties qui accompagnaient les soumissions non retenues. (C/T: 1207248) (Réf: 12-2)

Examples of Other Documents

CE-2011/4502
SOUMISSIONS «OS-27912» REJETÉES
RÉSOLU À L'UNANIMITÉ: que toutes les soumissions reçues portant le numéro «OS-27912» concernant l'acquisition et la mise en place d'un progiciel pour la gestion des heures et des horaires de travail (PPHT) soient et, par la présente, sont rejetées; que le Greffier ou la Greffière adjointe retourne les chèques ou garanties qui accompagnaient les soumissions. (Réf: 26-28)

Examples of Other Documents (cont.)

Immeuble - Aliénation
CE Mise en valeur du territoire et du patrimoine , Direction stratégies et transactions immobilières - 1074501006
Approuver un projet d'acte par lequel la Ville vend à Mark C. Moore, propriétaire des immeubles sis aux 3616-3652, rue Notre-Dame Ouest, un terrain situé dans l'arrondissement Le Sud-Ouest au sud-est de la rue Notre-Dame et au nord-est de la rue Bourget, constitué des lots 3 916 775 et 3 916 776 du cadastre du Québec, aux fins d'assemblage, d'une superficie totale d' environ 99,1 mètres carrés, pour le prix de 9 600 $, plus les taxes, si applicables, le tout sujet aux termes et conditions stipulés au projet d'acte.
Territoire(s) concerné(s) : Le Sud-Ouest District(s) : Saint-Henri - Petite-Bourgogne - Pointe-Saint-Charles

A Quick Perl Baseline

- Analysis of 36 documents
- 50 lines of perl
- Only documents with an amount and a company
- Only companies that end in Inc. (ignoring case)

A Quick Perl Baseline (cont.)

```
5   my@doc=<STDIN>;
6   chomp(@doc);
7   my$doc=join("␣",@doc); # one line, return characters transformed into spaces
8
9   # clean up header
10  if($doc=~ m/RÉSOLU À L'UNANIMITÉ\:/){
11  ␣␣␣␣$doc␣=~␣s/.*RÉSOLU␣À␣L'UNANIMITÉ://;
12  }
13
14  my@amounts = $doc =~
15    m/(\d?\d?(?:\s?|\.?)\d{3}(?:\s?|,?)(?:\d{3})?(?:\,\d{2})?\s?\$)/g;
16  if(!@amounts){ # no amount, bail-out
17      print "0\n"; exit;
18  }
```

A Quick Perl Baseline (cont.)

```
19  # has amount
20  if($#amounts){ # more than one amount, bail-out
21      print "0\n"; exit;
22  }
23  # got the amount, now find the company
24  my$amount = pop @amounts;
25
26  # is there a inc. ?
27  my@incs = split(/\sinc\./i, $doc);
28  if(!$#incs){ # no company, bail-out
29      print "0\n"; exit;
30  }
```

A Quick Perl Baseline (cont.)

```
32  # let's focus on the first one, we know $incs[0] ends in \sinc.
33  my$company = $incs[0];
34  # trim as much as possible
35  $company=~ s/.*firme//i;
36  $company=~s/.*à//i;
37  $company=~s/.*\spar\sle//i;
38  $company=~s/.*\sla\scompagnie\s//i;
```

A Quick Perl Baseline (cont.)

```
40  # reason is a crapshoot
41  my$reason="";
42  if($doc =~
43   m/((du\scontrat\sde)|(requis\spour)|(concernant)|
44  (\Q$amount\E\spour)|(\Q$company\E\s[Ii][Nn][Cc]\.\spour))/){
45      ($reason) = $doc =~ m/(?:(?:du\scontrat\sde)|(?:requis\spour)|
46  (?:concernant)|(?:\Q$amount\E\spour)|
47  (?:\Q$company\E\s[Ii][Nn][Cc]\.\spour))\s(.*)/;
48      # trim aggressively
49      $reason =~ s/(\,|\.|\;).*//;
50  }
51
52  print "1\t$amount\t$company␣inc.\t$reason\n";
```

Baseline Evaluation

- On the same 36 documents were it was programmed, at the document level by hand:

		System
Annotated	tp: 6	fn: 10
	fp: 0	tn: 20

  - precision = 6 / 6 = 1.0
  - recall = 6 / 16 = 0.375
- but partial reason in one and missing reason in another
  - 4 fn due to lack of amount, 6 due to lack of inc.

Baseline Evaluation (cont.)

- On 32 new documents, at the document level, by hand:

	System	
Annotated	tp: 5	fn: 9
	fp: 0	tn: 18

  - precision = 5 / 5 = 1.0
  - recall = 5 / 13 = 0.38
- but only 2 perfect, 1 wrong company and no reason, 1 no reason and 1 reason is whole doc
  - 2 fn due to lack of amount, 7 due to lack of inc.

Baseline Evaluation (cont.)

- More strict evaluation, using annotated CASes in XMI format
  - need to code the baseline CASes by hand
  - the whitespace conflation in perl destroys the offsets
  - See Baseline.md for details

	Development(36)			Test(32)		
	Prec	Rec	F	Prec	Rec	F
Amount	1.00	0.42	0.60	0.80	0.31	0.44
Company	0.83	0.18	0.29	0.80	0.21	0.33
Reason	0.40	0.10	0.15	0.33	0.06	0.10
Overall	0.77	0.21	0.33	0.69	0.18	0.29

Baseline Evaluation (cont.)

- https://github.com/IE4OpenData/ruta_testing_standalone
  - mvn package appassembler:assemble
  - ./target/appassembler/bin/ruta-evaluate
- Baseline.md

```
$ for t in Amount Company Reason; \
  do for s in 36 32; \
    do echo $t $s; \
    /path/to/ruta_testing_standalone/target/appassembler/bin/ruta-evaluate \
      --gold data/gold$s --eval data/baseline$s \
      --include org.ie4opendata.octroy.$t \
      --typesystem ./src/main/resources/org/ie4opendata/octroy/octroy_eval_ts.xml; \
    done;
   done
```

What is to come

- Entity Detection and Linking
- Rule-based IE
- Statistical IE
- Integration / Research directions

Class 2: Entity Detection and Linking

- Named Entity Recognition
  - Lists of words
  - Regular Expressions
  - Basic ML
  - Beginning / Inside / Out annotation
- Entity Linking
  - Word Sense Disambiguation
  - Context Search Engine
  - DBpedia Spotlight

Class 3: Rule-Based IE

- RE-based rules
- Anchor-based rules
- Rule-learning
- Apache RuTA

Class 4: Statistical IE

- Sequence tagging
- HMMs / MEMMs / CRFs
- ClearTk

Class 5: Integration

- Integration
- Annotation

Class 6: Research

- Research directions
  - Open IE
  - Cross-document IE
  - Multilingual IE
  - Multimedia IE
  - Adaptation for IE
  - Deep Learning

Course Evaluation

- To pass this course I am asking you to take the class example and either
  - extended / modify it over the existing data OR
  - Apply it to a different dataset
- The output is a one-page report we will discuss the last class

Handout

1. Chapter 22 -- Information Extraction from Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing: Second Edition, Daniel Jurafsky & James H. Martin. Draft of October 15, 2007.
2. Cunningham H (2006), Information Extraction, Automatic. In: Keith Brown, (Editor-in-Chief) Encyclopedia of Language & Linguistics, Second Edition, volume 5, pp. 665-677.
3. Chapter 7 -- Extracting Information from Text Natural Language Processing with Python, by Steven Bird, Ewan Klein and Edward Loper, Copyright (C) 2014 the authors.
4. Chapter 2 -- UIMA Conceptual Overview UIMA Documentation Overview by The Apache UIMA Development Community. Version 2.8.1 -- Copyright (C) 2006 The Apache Software Foundation
5. UIMA RuTA tutorial by Peter Klügl at GSCL 2013.

Daily Feedback

- At the end of each class, take a moment to write on a piece of paper:
  - Any questions or comments you might have
  - A brief summary of the topics of the day seen in class
- Put your name in the paper to track attendance
- Your questions and comments will be addressed the next class
- This is mandatory