



# CSCE 771: Computer Processing of Natural Language

## Lecture 17: Entity Extraction, Linking

---

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

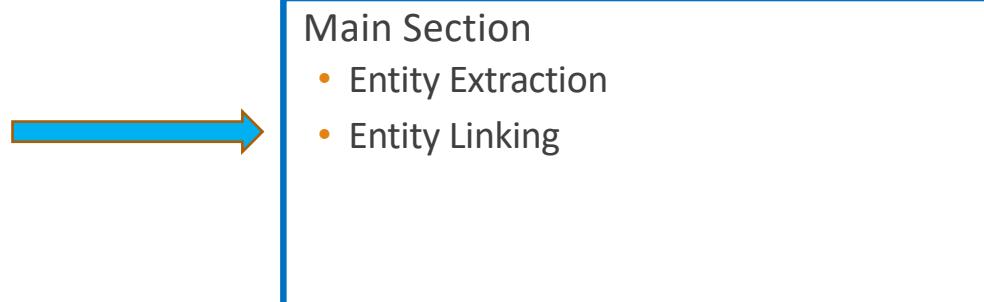
18<sup>TH</sup> OCTOBER, 2022

***Carolinian Creed: “I will practice personal and academic integrity.”***

# Organization of Lecture 17

---

- Opening Segment
  - Announcements
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 18



# Recent Classes

---

Oct 11 (Tu)	Guest Lecture – Dr. Amitava Das: Using lang models to solve NLP tasks
Oct 13 (Th)	
Oct 18 (Tu)	Entity extraction, linking
Oct 20 (Th)	Events extraction, spatio-temporal analysis
Oct 25 (Tu)	Topic Analysis
Oct 27 (Th)	PROJ REVIEW
Nov 1 (Tu)	NLP Task: Sentiment
Nov 3 (Th)	NLP Task: Summarization

## Review of Lecture 16

- Transformers
  - Motivation and history
  - Attention types
- Paper readings
  - Hang Li, [Language Models: Past, Present, and Future](#), Communications of the ACM, July 2022, Vol. 65 No. 7, Pages 56-63 10.1145/3490443
  - [A Primer in BERTology: What We Know About How BERT Works](#) (Rogers et al., TACL 2020)

# Announcements

---

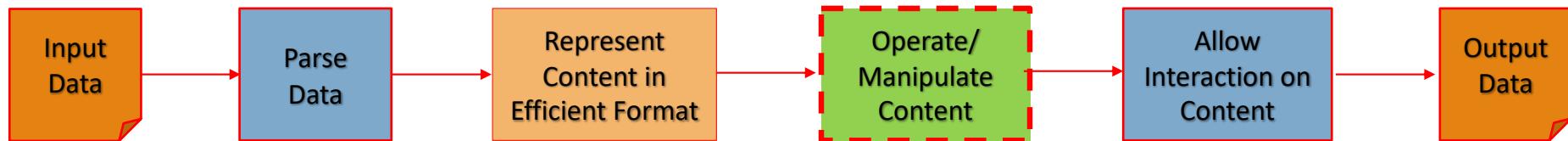
- Participate in mid-course survey: <https://forms.gle/JNMPvUMG2pvs1r5U9> (optional, due by Oct 20, 2022)
- Quizzes moved to focus on NLP tasks and exploration of LLMs for tasks; projects

# Main Lecture

---

# Methods to Extract Content

---



# What is an Entity?

---

- Definition
  - Oxford: “a thing with distinct and independent existence”
  - Practical: Any mention in text of interest
- Types
  - Physical: Person, animal, mountain
  - Abstract: Emotion, nation, money
- Heuristic: Entities are often nouns

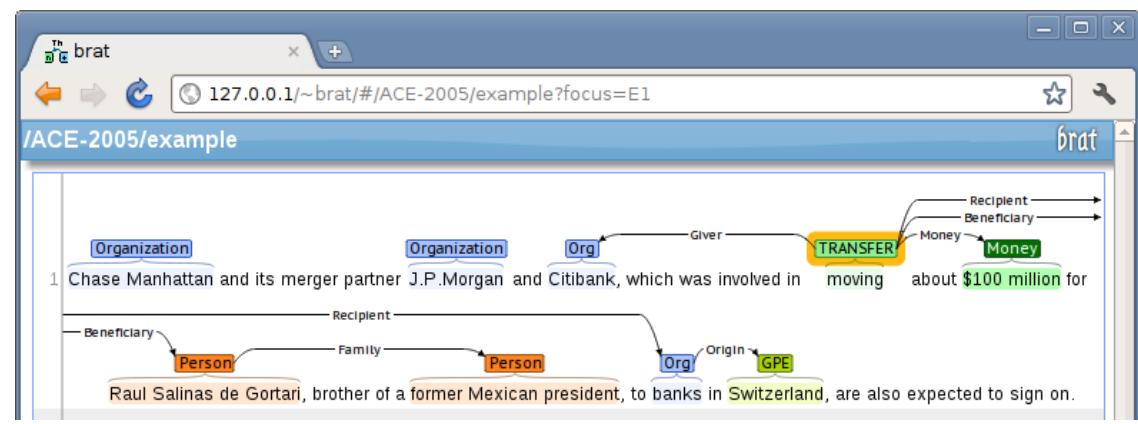
The **Nobel Peace Prize** is one of the five **Nobel Prizes** established by the **will** of Swedish industrialist, inventor, and armaments manufacturer **Alfred Nobel**, along with the prizes in **Chemistry**, **Physics**, **Physiology or Medicine**, and **Literature**

Credit: From Wikipedia

# Entity Extraction Methods

- Regular expression: find patterns in content
  - Why: if pattern known, easy, fast and cheap to implement
  - Why not: pattern has to be known
- Manual annotation: tag entities and store in a repository; runtime - match in content and retrieve tags
  - Tool: BRAT - <https://brat.nlplab.org/introduction.html>
  - Why: use information when available
  - Why not: cost of annotation is high, time-consuming

*Also called: Entity identification, entity chunking, Named entity recognition (NER)*



Annotate entity types and relationships; Credit: <https://brat.nlplab.org/introduction.html>

Reference: <https://lionbridge.ai/articles/the-essential-guide-to-entity-extraction/>

# Entity Extraction – Methods Continued

---

- Learning based – many varieties
  - Why: Pretrained available, domain-specific models, alignment with standards
  - Why not: needs large compute resources, may not be explainable

Reference: <https://lionbridge.ai/articles/the-essential-guide-to-entity-extraction/>

# Which Learning-based Method

---

- Conditional random field (CRF): learn probability of entities based on defined features over inputs
  - Requires labeled data about text and entities, needs features, learns entity labels
  - Articles: <https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html#let-s-use-conll-2002-data-to-build-a-ner-system>; <https://www.depends-on-the-definition.com/named-entity-recognition-conditional-random-fields-python/>
- LSTM-based: predict labels (entities) over text sequences.
  - Requires labeled data about text and entities, models forward and backward neighborhood, learns entity labels
  - Blog: <https://www.depends-on-the-definition.com/named-entity-recognition-with-residual-lstm-and-elmo/>
- Deep learning based models
  - A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, [Vikas Yadav, Steven Bethard](#), ACL 2018 <https://www.aclweb.org/anthology/C18-1182.pdf>

# Benchmarks

- <https://paperswithcode.com/task/named-entity-recognition-ner/codeless>

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	CoNLL 2003 (English)	🏆 CNN Large + fine-tune	Cloze-driven Pretraining of Self-attention Networks			<a href="#">See all</a>
	Ontonotes v5 (English)	🏆 BERT-MRC+DSC	Dice Loss for Data-imbalanced NLP Tasks			<a href="#">See all</a>
	ACE 2005	🏆 BERT-MRC	A Unified MRC Framework for Named Entity Recognition			<a href="#">See all</a>
	GENIA	🏆 BERT-MRC	A Unified MRC Framework for Named Entity Recognition			<a href="#">See all</a>
	CoNLL++	🏆 CrossWeigh + Pooled Flair	CrossWeigh: Training Named Entity Tagger from Imperfect Annotations			<a href="#">See all</a>
	Long-tail emerging entities	🏆 Flair embeddings	Contextual String Embeddings for Sequence Labeling			<a href="#">See all</a>
	BC5CDR	🏆 NER+PA+RL (PubMed)	Reinforcement-based denoising of distantly supervised NER with partial annotation			<a href="#">See all</a>
	JNLPBA	🏆 BioBERT	BioBERT: a pre-trained biomedical language representation model for biomedical text mining			<a href="#">See all</a>
	SciERC	🏆 SpERT	Span-based Joint Entity and Relation Extraction with Transformer Pre-training			<a href="#">See all</a>

# Annotation of Entities for Interchange

---

- IOB stands for inside-outside-beginning
- Standoff format

## Named Entity types

- person names (PER),
- organizations (ORG),
- locations (LOC) and
- Times
- Quantities
- Miscellaneous names (MISC)

CONLL shared tasks

2002: <https://www.aclweb.org/anthology/W02-2024/>

2003: <https://www.aclweb.org/anthology/W03-0419.pdf>

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Source: <https://lionbridge.ai/articles/the-essential-guide-to-entity-extraction/>

# BioMedical Entity Extraction

---

- HunFlair - <https://github.com/flairNLP/flair/blob/master/resources/docs/HUNFLAIR.md>
- Has learned models for genes/proteins, chemicals, diseases, species and cell lines.

**Input:** "Behavioral abnormalities in the Fmr1 KO2 Mouse Model of Fragile X Syndrome"

Span [1,2]: "Behavioral abnormalities" [- Labels: Disease (0.6736)]

Span [10,11,12]: "Fragile X Syndrome" [- Labels: Disease (0.99)]

Span [5]: "Fmr1" [- Labels: Gene (0.838)]

Span [7]: "Mouse" [- Labels: Species (0.9979)]

**Credit:** <https://github.com/flairNLP/flair/blob/master/resources/docs/HUNFLAIR.md#quick-start>

# Case Study: Information Extraction and Linking for Business Documents

---

## References:

- Elad Fein, Natalia Razinkov, Shlomit Shachor, Pietro Mazzoleni, Sweefen Goh, Richard Goodwin, Manisha Bhandar, Shyh-Kwei Chen, Juhnyoung Lee, Vibha Singhal Sinha, Senthil Mani, Debdoot Mukherjee, Biplav Srivastava, and Pankaj Dhoolia. 2011. Using MATCON to generate CASE tools that guide deployment of pre-packaged applications. In Proceedings of the 33rd International Conference on Software Engineering (ICSE '11). ACM, New York, NY, USA, 1016-1018. DOI:<https://doi.org/10.1145/1985793.1985981>
- Biplav Srivastava, Debdoot Mukherjee, Rema Ananthanarayanan and Vibha Sinha, Model Extraction to Model-based Reuse of Enterprise Documents, In ACM International Conference on Management of Data (COMAD 2010), Nagpur, India, 8-10 December 2010. [Area:Content Harvesting-Model extraction]
- Debdoot Mukherjee ; Senthil Mani ; Vibha Singhal Sinha ; Rema Ananthanarayanan ; Biplav Srivastava, Pankaj Dhoolia ; Prahlad Chowdhury, "AHA: Asset Harvester Assistant," 2010 IEEE International Conference on Services Computing, Miami, FL, 2010, pp. 425-432. [Area:Content Harvesting-Services]
- Biplav Srivastava, Yuan-chi Chang, Business Insight from Collection of Unstructured Formatted Documents with IBM Content Harvester, In ACM International Conference on Management of Data (COMAD 2009), Mysore, India, 9-12 December 2009. [Area:BusinessProcess-DocHarvesting]
- Biplav Srivastava, Debdoot Mukherjee, Organizing Documented Processes, In IEEE International Conference on Services Computing (SCC 2009), Bangalore, India, September 21-25, 2009.[Area:BusinessProcess-TextFlowAnalyses]

# Business Setting

---

- **Setting:** Enterprise Resource Planning Projects
  - Off-the-shelf software to manage common business functions (e.g. Finance, Supply Chain)
  - Businesses buy these software and then engage service providers to tailor them
  - System Integration Market Worth \$582.5 Billion By 2025,  
<https://www.grandviewresearch.com/press-release/global-system-integration-market>
  - AMR Research estimates that spending on consulting, integration and support for packaged application services was \$103B in 2007, and expected to reach \$174B by 2012.
- Business Processes are captured in large numbers and in multiple representations
  - Typically over 100 business processes per engagement
  - Flow Diagrams: Visio, PowerPoint
  - Text Documents: Word, Excel



# Diverse Extraction Units

- An extraction unit is a contiguous sequence of content from a document that can be assigned a tag.
  - Plain text, list, and rich text are different types of extraction units possible

Process	<b>Procurement Master Data</b>
Team	<b>Plan to Stock</b>
Approver	<b>Jan Meier</b>
Version	<b>01</b>

**User Requirements Specification Completed**

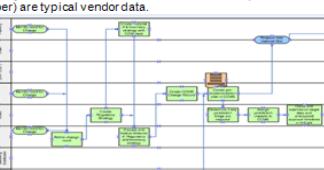
## Section 1: Definition

### Description

MM Purchasing processes the following types of data:

**Material Master Data**  
 Details on materials that an enterprise procures externally or produces in-house. The unit of measure and the description of a material are examples of the data stored in a material master record. Other SAP Logistics components also access the material data.

**Vendor Master Data**  
 Information about external/internal suppliers (creditors). The vendor's name and address, the currency the vendor uses, and the vendor number (stored in the SAP system as an account number) are typical vendor data.



Purchasing Master Data, such as the following:

- Purchasing Info Record**  
 The info record establishes the link between material and vendor, thus facilitating the process of selecting a source. For example, the info record shows the unit of measure used for ordering from the vendor, and indicates vendor price changes affecting the material over a period of time.
- Quota Arrangement**  
 The quota arrangement specifies which portion of the total requirement of a material over a certain period is to be assigned to particular vendors on the basis of quotas.
- Approved Manufacturers Part List (AMPL)**  
 The AMPL specifies:
  - Whether a material can be produced in-house or by an external vendor

**Simple string of words**

**Rich text content**

- The period during which the MPN material may be procured
- Whether it must belong to a certain revision level
- Whether it may only be procured from a certain manufacturer plan
- Whether it is blocked and, if so, why

**Triggers**

New Product Introduction  
New Vendor  
Request to Inventory a Material  
Change in Supplier Status  
Change in Material Status

**Dependencies/Requirements**

- Default lot status at receipt
- Material identification (commercial vs. clinical or MFG stages)

**Inputs**

- Master Data Details
  - Material
  - Vendor
  - Info Record
  - Source List
  - Quota Arrangement
  - Delivery Address
  - Bill to Address

**Supplier**

- QA
- PO
- Planner
- Planner/Buyer
- Supplier
- Finance

**Outputs**

Master Data Record Created for the views and Customers

Planner/Buyers, Planners, Manufacturing, M Steps

Final description of each of the boxes in the process

**Action**

- Maintain Material Master
- Maintain Vendor Master

**Simple record instance model.**

**Complex entity record in the form of a particular column in turn in the entity**

# Key Information Elements

## Process Specific Artifacts

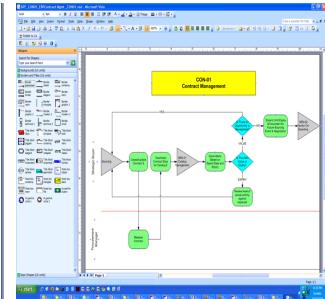
## Scenario

# Process

## Process Step

## Inputs, Outputs

ERP Project		
Process Definition Document		
System Upgraded New System Presented		
<b>Customers</b>		
	<ul style="list-style-type: none"> <li>• Applications/Member</li> <li>• Manager</li> <li>• DAF</li> <li>• Help Desk</li> <li>• Information Center</li> <li>• Internal Audit</li> <li>• Project Manager</li> <li>• Help Desk</li> <li>• Manager</li> </ul>	
<b>Suppliers</b>		
23.1	<p>Demand (what's it like in the Americas)</p> <p>If no, complete purchasing request notes (including system costs) and forward information to HNC to receive quote.</p>	Manager Manager Manager Manager
23.2	<p>Send purchasing request notes to HNC (US, Local HR, Gen 2.3 &amp; 2.4, and 2.5).</p> <p>For major serial reordering link from HNC and generate work orders (Gen 2.3 &amp; 2.4).</p>	Manager Manager Manager Manager
23.3	<p>Approve and release required purchase - complete purchase requisition (Gen 2.3 &amp; 2.4).</p> <p>Complete New Purchase process</p>	Other Manager



## Non-Process Business Artifacts

## Requirement

## Use-case

Gap

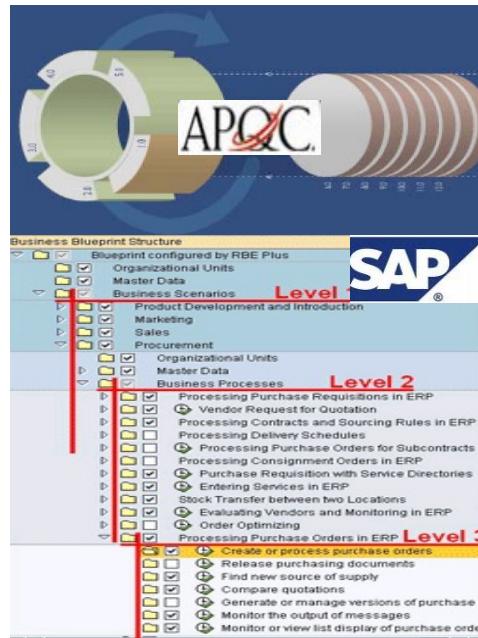
KPI



# Business Process Hierarchies

## ➡ Industry Specific

 Cross Industry



## Data, Data Everywhere... Nor Any Drop to Use!!

---

- Design information on business artifacts implemented in engagements are locked in documents
  - Need to turn them into reusable assets
  - Retrieve information into a model based format
- Enterprise asset repositories are not well organized
  - Essentially, a dump of unlinked process documentation in different formats
  - No meta-data available against silos of documents
- Inconsistencies in process data
  - Multiple teams are responsible for various aspects of process design

## Research Theme

---

- Establish an effective framework for organizing ***design-level documentation*** on business processes and ***linked business artifacts*** in order to:
  - Boost information reuse across engagements
  - Maintain coherence in enterprise process repositories
  - Reduce costs and improve quality in business transformation exercises
- Effective reuse of process information from past engagements will yield great benefits
  - Conventional document management systems are not capable of providing a process-centric view of information
  - How to search for the most effective business artifacts in the current “process” context?

# Summary of Prior Art

---

- More than a decade of literature exists in information extraction from Web data
- *Wrappers* are programs to extract information from web pages
- Broadly 3 categories of information extractors:
  - **Rule-based:** Extraction patterns are specified using some specialized notation
  - **Supervised:** Wrappers are inferred based on a set of examples provided by the user
  - **Unsupervised:** Infer the content extraction patterns based on a collection of documents
- **AHA Approach:** Unsupervised followed by Manual Review
  - Dependent on Word's document model
  - Another approach tried called **Content Harvester**, which was independent of tool-specific model

## Related Literature

- Work in measuring similarity (diagnosing differences) in business process models
  - e.g., Ehrig et al (APCCM '07), Dijkman (BPM '08), Van der Aalst et al (BPM '06)
  - Compares flow models in structured formats viz. Petri net, EPC, YAWL
  - Linguistic, semantic and structural dimensions of comparing process elements
- Extensive literature in Process Mining from execution logs
  - ProM framework
- Research on choosing an appropriate granularity of process model reuse
  - Holschke et al (BPM '09), Mendling et al (BPM '08)
- Extraction and management of useful process variants (Sadiq BPM '06)
- Traditional methods in legacy text mining and organization
  - But they do not specifically focus on process information
- ***No known effort to target design level process information with linkage to business artifacts of interest viz. requirements, KPIs, use-cases***

# Diverse Extraction Units

- An extraction unit is a contiguous sequence of content from a document that can be assigned a tag.
  - Plain text, list, and rich text are different types of extraction units possible

Process	<b>Procurement Master Data</b>
Team	<b>Plan to Stock</b>
Approver	<b>Tan Meier</b>
Version	<b>Q1</b>

**User Requirements Specification Completed**

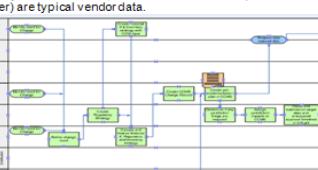
## Section 1: Definition

### Description

MM Purchasing processes the following types of data:

**Material Master Data**  
 Details on materials that an enterprise procures externally or produces in-house. The unit of measure and the description of a material are examples of the data stored in a material master record. Other SAP Logistics components also access the material data.

**Vendor Master Data**  
 Information about external/internal suppliers (creditors). The vendor's name and address, the currency the vendor uses, and the vendor number (stored in the SAP system as an account number) are typical vendor data.



```

graph TD
    PO[Purchase Order] --> MM[MM Purchasing]
    MM --> MMD[Material Master Data]
    MM --> VMD[Vendor Master Data]
    MMD --> MMR[Material Master Record]
    VMD --> VMR[Vendor Master Record]
    MMR --> MRP[Material Requirements Planning]
    VMR --> QAR[Quota Arrangement]
    MRP --> AMPL[Approved Manufacturers Part List]
    QAR --> AMPL
  
```

Purchasing Master Data, such as the following:

- Purchasing Info Record**  
 The info record establishes the link between material and vendor, thus facilitating the process of selecting a source. For example, the info record shows the unit of measure used for ordering from the vendor, and indicates vendor price changes affecting the material over a period of time.
- Quota Arrangement**  
 The quota arrangement specifies which portion of the total requirement of a material over a certain period is to be assigned to particular vendors on the basis of quotes.
- Approved Manufacturers Part List (AMPL)**  
 The AMPL specifies:
  - Whether a material can be produced in-house or must be purchased

**Rich text content**

- Whether a material can be produced in-house or must be purchased

- The period during which the MPN material may be procured
- Whether it must belong to a certain revision level
- Whether it may only be procured from a certain manufacturer plan
- Whether it is blocked and, if so, why

**Triggers**

New Product Introduction  
New Vendor  
Request to Inventory a Material  
Change in Supplier Status  
Change in Material Status

**Dependencies/Requirements**

- Default lot status at receipt
- Material identification (commercial vs. clinical or MFG stages)

**Inputs**

- Master Data Details
  - Material
  - Version
  - Info Record
  - Source List
  - Quota Arrangement
  - Delivery Address
  - Bill to Address

**Supplier**

- QA
- PO
- Planner
- Planner/Buyer
- Supplier
- Finance

**Outputs**

Master Data Record Created for the views and Customers

Planner/Buyers, Planners, Manufacturing, M Steps

Final description of each of the boxes in the process

**Action**

- Maintain Material Master
- Maintain Vendor Master

**Simple record instance model.**

**Complex entity record in the form of a particular column in turn in the entity**

# Process Information Extraction - Text

- Utilize semi-structured nature of data
- Extract content segments present in a document collection, which can map to some process semantics
- Seek an appropriate tag (preferably from a pre-defined meta model) from the user
- Utilize layout of content segments in the document to establish cardinality and relations between various pieces of flat tagged content

The diagram illustrates the process of extracting and tagging content from a document. On the left, a screenshot of Microsoft Word shows a document titled "Project\_Structure.xls" containing sections like "User Requirements Specification Completed", "Section 1: Definition", and "Section 2: Requirements". Annotations include a blue arrow pointing to the word "Extract" and another blue arrow pointing to the word "Tag". On the right, a screenshot of Microsoft Excel shows a table titled "Section[1\$]" with columns A through E. The table contains rows of content segments, each with a "URL" column linking to the document. The "Tag" column is highlighted in green, indicating the extracted tags. Below the table, a snippet of XML code shows how these tags are mapped to specific steps and roles in a process flow.

	A	B	C	D	E
1 Identifier	URL	Navigate to Word D Entity Tag			Attribute
2 Document		<a href="#">View Content</a>			
3 Table[1\$]		<a href="#">View Content</a>			
<b>7 Section[1\$]</b>		<a href="#">View Content</a>			
8 Section[Section 1: Definition]		<a href="#">View Content</a>			
9 Section[Customers]	file:///C:/AHA data\am\	<a href="#">View Content</a>			
10 Table[Steps]		<a href="#">View Content</a>			
11 Column[1]		<a href="#">View Content</a>			
12 Column[Action]	file:///C:/AHA data\am\	<a href="#">View Content</a>			Step
13 Column[By Whom ]		<a href="#">View Content</a>			Role
14 Column[Manual or System]		<a href="#">View Content</a>			name
15 Column[Users]	file:///C:/AHA data\am\	<a href="#">View Content</a>			Customer
16 Column[Dependencies/Requirements]		<a href="#">View Content</a>			name
17 Column[Inputs]	file:///C:/AHA data\am\	<a href="#">View Content</a>			Input
18 Column[Outputs]		<a href="#">View Content</a>			name
19 Column[Design Decisions]	file:///C:/AHA data\am\	<a href="#">View Content</a>			Variant
20 Column[keyDesign]		<a href="#">View Content</a>			keyDesign
21 Section[Refacing Systems]		<a href="#">View Content</a>			
22 Section[Current Process Attributes]		<a href="#">View Content</a>			

```

<step name="Analyze Client Needs: Sourcing manager will meet with client lead to identify supplier pain points, potential spend targets, <role name="Sourcing Manager">">
<step>
<step name="Analyze Existing Supplier Capabilities: Sourcing manager will identify existing suppliers in relevant categories of spend at List of suppliers in general area of spend to be maintained in SAP, <role name="Sourcing Manager">">
<step>
<step name="Analyze Historical Spend: Sourcing manager will direct analytical group to pull spend data in relevant categories of spend, <role name="Sourcing Analytics">">
<step>
<step name="Conduct Market Scan: Sourcing analytical group will utilize external resources to identify alternative suppliers as well as, <role name="Sourcing Analytics">">
<step>
<step name="Identify Gaps Opportunities: Sourcing manager and client lead will create a list of potential opportunities for sourcing with, <role name="Sourcing Manager">">
<step>
<step name="Prioritize Opportunities: Client lead will present opportunities to management to determine which sourcing opportunities client lead will work with FP&A, <role name="Client Lead">">
<step>
<step name="Forecast Savings: Sourcing manager will create an estimate of timing, spend and savings potential for projects scheduled, <role name="Sourcing Manager">">
<step>

```

# Entity Matching

---

- Matching data instances that refer to the same real-world entity
- Also called
  - entity resolution
  - reference reconciliation
  - deduplication

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to \$2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Credit: Jurafsky and Martin

## Sources:

- Magellan: <https://sites.google.com/site/anhaiingroup/projects/magellan>
- Summary: <https://cacm.acm.org/magazines/2020/8/246370-technical-perspective-entity-matching-with-magellan/fulltext>
- Paper: <https://cacm.acm.org/magazines/2020/8/246373-magellan/fulltext>

# Example and Terms

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to \$2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

**Mentions:** linguistic term in text referring to real world/domain of discourse entity - Victoria Chen, her

**Referent:** Entity in the domain - Victoria Chen (the person)

**Corefer:** One or more terms referring to a referent –

Victoria Chen, her, the 38-year-old, she

**Anaphora:** One or more mentions referring to a previously introduced referent –

her, the 38-year-old, she

**Coreference chain or cluster:** set of corefering expressions

## *Coreference chains in example*

1. {*Victoria Chen*, *her*, *the 38-year-old*, *She*}
2. {*Megabucks Banking*, *the company*, *Megabucks*}
3. {*her pay*}
4. {*Lotsabucks*}

Source: Jurafsky and Martin

# Famous Example: Winograd Schema Challenge

Three parts:

- A sentence or brief discourse that contains the following:
  - Two **noun phrases** of the same semantic class (male, female, inanimate, or group of objects or people),
  - An ambiguous **pronoun** that may refer to either of the above noun phrases, and
  - A special word and alternate word, such that if the special word is replaced with the alternate word, the natural resolution of the pronoun changes.
- A question asking the identity of the ambiguous pronoun, and
- Two answer choices corresponding to the noun phrases in question.

The city council denied the demonstrators a permit because

- a. they feared violence.
- b. they advocated violence.

They refers to:

- City council or demonstrators ?

Sources: Jurafsky and Martin, [https://en.wikipedia.org/wiki/Winograd\\_Schema\\_Challenge](https://en.wikipedia.org/wiki/Winograd_Schema_Challenge)

# WinoBias - Bias in Windograd Setting

- Sentence contain two mentions corresponding to stereotypically-male and stereotypically-female occupations and a gendered pronoun that must be linked to one of them.
  - The sentence cannot be disambiguated by the gender of the pronoun.
- A coreference system is biased if
  - more accurate at linking pronouns consistent with gender stereotypical occupations (e.g., **him with physician**) than linking pronouns inconsistent with gender-stereotypical occupations (e.g., **her with physician**).

The secretary called the physician<sub>i</sub> and told him<sub>i</sub> about a new patient [pro-stereotypical]  
The secretary called the physician<sub>i</sub> and told her<sub>i</sub> about a new patient [anti-stereotypical]

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. In NAACL HLT 2018.

**Result:** Systems of all architectures (rule-based, feature-based machine learned, and end-to-end-neural) all show significant bias, performing on average 21 F1 points worse in the anti-stereotypical cases

# Magellan System

- A platform to use and create entity matching techniques and building blocks. Components: data cleaning, visualization, blocking and matching
- Two entity matching tools
  - (1) PyMatcher is an entity matching tool that in the PyData ecosystem.
  - (2) CloudMatcher is a cloud-based entity matching tool that is part of the Amazon Web Services ecosystem.

Problem owner	Problem type	Notable result	In production?	Team	Other
<b>Walmart</b>	Debug a system in production that matches products	Improved recall by 34%, reduced precision by 0.65%	Yes	1 student, 1 employee	Funding
<b>Recruit holdings</b>	Matching names of stores, companies, properties	Reported 98.9% accuracy on matching 10K store names	Yes	Multiple employees	Press release
<b>Johnson controls Marshfield clinic</b>	Matching suppliers Matching drugs	Precision and recall in 96–100% 99.2% precision and 95.3% recall	Unknown Yes	1 student 1 student, 1 employee 2 students	Funding Paper
<b>Economics (UW)</b>	Matching grants. Build a better EM pipeline	Precision in [96.7%, 98.8%], recall in [94.2%, 97.1%] A system in production achieves 100% precision, recall in [65.1%, 71.8%]	Not yet	2 students	Paper published, funding from UW
<b>Land use (UW)</b>	Matching cattle ranches. Build a better EM pipeline	Precision in [89.7%, 99.0%], recall in [79.2%, 92.2%] A system in production achieves precision in [94.9%, 100%], recall in [29.4%, 46.6%]	Yes	1 student, 1 programmer, 2 staff persons	Paper planned, funding from UW
<b>Biomedicine (UW)</b>	Matching ontology terms	metasra.biostat.wisc.edu	Yes	1 student	Paper published
<b>Limnology (UW)</b>	Matching table attributes	High-value data sets created from multiple data sets	Yes	2 students	Funding from UW

Sources:

- Magellan: <https://sites.google.com/site/anhaiingroup/projects/magellan>
- Summary: <https://cacm.acm.org/magazines/2020/8/246370-technical-perspective-entity-matching-with-magellan/fulltext>
- Paper: <https://cacm.acm.org/magazines/2020/8/246373-magellan/fulltext>

# Entity Linking

- Matching entity to a real-world individual
- Also called entity resolution (as per Jurafsky and Martin)
- How do we know real-world individuals ?
  - Mapping to an authoritative ontology
    - Wikipedia entry (**wikification**)
    - Medical ontology - MESH
  - List of entities – Gazetteer

## Wikipedia Example: White House

[https://en.wikipedia.org/wiki/White\\_House](https://en.wikipedia.org/wiki/White_House)

Categories: [White House](#) | [Houses completed in 1800](#) | [Buildings of the United States government in Washington, D.C.](#)  
[Houses on the National Register of Historic Places in Washington, D.C.](#) | [National Historic Landmarks in Washington, D.C.](#) | [Presidential residences in the United States](#)  
[Presidential residences](#) | [Palaces in the United States](#) | [Rebuilt buildings and structures in the United States](#) | [Reportedly haunted locations in Washington, D.C.](#)  
[Federal architecture in Washington, D.C.](#) | [Neoclassical architecture in Washington, D.C.](#) | [Presidential museums in Washington, D.C.](#) | [Historic house museums in Washington, D.C.](#)  
[Neoclassical palaces](#) | [Buildings and structures in the United States destroyed by arson](#) | [1800 establishments in Washington, D.C.](#) | [James Hoban buildings](#)  
[Burned houses in the United States](#) | [Presidential homes in the United States](#) | [Northwest \(Washington, D.C.\)](#)

## Contents [hide]

<a href="#">1 Early history</a>
<a href="#">1.1 1789–1800</a>
<a href="#">1.2 Architectural competition</a>
<a href="#">1.3 Design influences</a>
<a href="#">1.4 Construction</a>
<a href="#">1.5 Architectural description</a>
<a href="#">1.6 Naming conventions</a>
<a href="#">2 Evolution of the White House</a>
<a href="#">2.1 Early use, the 1814 fire, and rebuilding</a>
<a href="#">2.2 Overcrowding and building the West Wing</a>
<a href="#">2.3 Truman reconstruction</a>
<a href="#">2.4 Jacqueline Kennedy restoration</a>
<a href="#">3 The White House since the Kennedy restoration</a>
<a href="#">3.1 Layout and amenities</a>
<a href="#">3.2 Executive Residence</a>
<a href="#">3.3 West Wing</a>
<a href="#">3.4 East Wing</a>
<a href="#">3.5 Grounds</a>
<a href="#">4 Public access and security</a>
<a href="#">4.1 Historical accessibility</a>
<a href="#">4.2 Aviation incidents</a>
<a href="#">4.3 Closure of Pennsylvania Avenue</a>
<a href="#">4.4 Protection</a>
<a href="#">5 See also</a>
<a href="#">6 Notes</a>
<a href="#">7 References</a>
<a href="#">8 Further reading</a>
<a href="#">9 External links</a>

## Empirical Evaluation — Results

---

Reading material

Large-Scale Named Entity Disambiguation  
Based on Wikipedia Data, Silviu Cucerzan,

Proceedings of the 2007 Joint Conference on Empirical Methods in  
Natural Language Processing and Computational Natural Language  
Learning, pp. 708–716, Prague, June 2007.

# Matching Entity to Wikipedia Entries

Document context:

- Mentions
- Named entities
- Coreferences
- disambiguations

Wikipedia Example: White House

[https://en.wikipedia.org/wiki/White\\_House](https://en.wikipedia.org/wiki/White_House)

## Contents [hide]

1	<a href="#">Early history</a>
1.1	<a href="#">1789–1800</a>
1.2	<a href="#">Architectural competition</a>
1.3	<a href="#">Design influences</a>
1.4	<a href="#">Construction</a>
1.5	<a href="#">Architectural description</a>
1.6	<a href="#">Naming conventions</a>
2	<a href="#">Evolution of the White House</a>
2.1	<a href="#">Early use, the 1814 fire, and rebuilding</a>
2.2	<a href="#">Overcrowding and building the West Wing</a>
2.3	<a href="#">Truman reconstruction</a>
2.4	<a href="#">Jacqueline Kennedy restoration</a>
3	<a href="#">The White House since the Kennedy restoration</a>
3.1	<a href="#">Layout and amenities</a>
3.2	<a href="#">Executive Residence</a>
3.3	<a href="#">West Wing</a>
3.4	<a href="#">East Wing</a>
3.5	<a href="#">Grounds</a>
4	<a href="#">Public access and security</a>
4.1	<a href="#">Historical accessibility</a>
4.2	<a href="#">Aviation incidents</a>
4.3	<a href="#">Closure of Pennsylvania Avenue</a>
4.4	<a href="#">Protection</a>
5	<a href="#">See also</a>
6	<a href="#">Notes</a>
7	<a href="#">References</a>
8	<a href="#">Further reading</a>
9	<a href="#">External links</a>

Wikipedia context:

- From url,
- Disambiguations,
- Content hierarchies
- categories

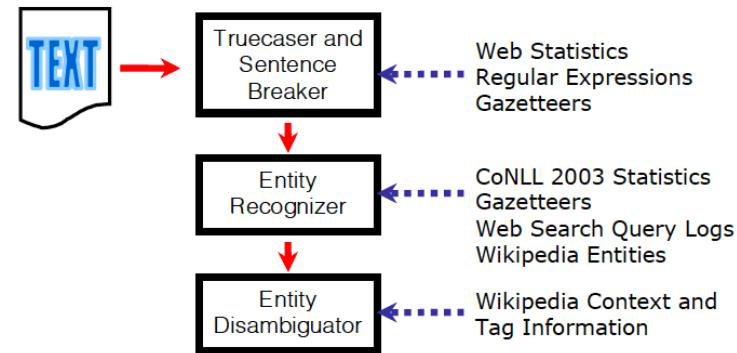
Categories: [White House](#) | [Houses completed in 1800](#) | [Buildings of the United States government in Washington, D.C.](#)  
[Houses on the National Register of Historic Places in Washington, D.C.](#) | [National Historic Landmarks in Washington, D.C.](#) | [Presidential residences in the United States](#)  
[Presidential residences](#) | [Palaces in the United States](#) | [Rebuilt buildings and structures in the United States](#) | [Reportedly haunted locations in Washington, D.C.](#)  
[Federal architecture in Washington, D.C.](#) | [Neoclassical architecture in Washington, D.C.](#) | [Presidential museums in Washington, D.C.](#) | [Historic house museums in Washington, D.C.](#)  
[Neoclassical palaces](#) | [Buildings and structures in the United States destroyed by arson](#) | [1800 establishments in Washington, D.C.](#) | [James Hoban buildings](#)  
[Burned houses in the United States](#) | [Presidential homes in the United States](#) | [Northwest \(Washington, D.C.\)](#)

# Extracting and Linking Entities with Authoritative Knowledge Source

Key idea: use contextual and category information from Wikipedia

Approach:

- maximize agreement of context between
  - contextual information extracted from Wikipedia
  - context of a document, as well as the
- agreement among the category tags associated with the candidate entities



# Entity Linking

- State of art and datasets
- Datasets
  - OntoNotes
  - AIDA-CoNLL
  - TAC KBP 2010

https://paperswithcode.com/task/entity-linking

### Benchmarks

TREND	DATASET	BEST METHOD	PAPER TITLE	PDF
	AIDA-CoNLL	🏆 Flair/rel-norm	REL: An Entity Linker Standing on the Shoulders of Giants	
	WebQSP-WD	🏆 VCG	Mixing Context Granularities for Improved Entity Linking on Question Answering Data across Entity Categories	
	CoNLL-Aida	🏆 RELIC + CoNLL-Aida tuning	Learning Cross-Context Entity Representations from Text	
	TAC-KBP 2010	🏆 RELIC + CoNLL-Aida tuning	Learning Cross-Context Entity Representations from Text	
	FIGER	🏆 ERNIE	ERNIE: Enhanced Language Representation with Informative Entities	
	MSNBC	🏆 E2E	End-to-End Neural Entity Linking	
	OKE-2015	🏆 E2E	End-to-End Neural Entity Linking	
	OKE-2016	🏆 E2E	End-to-End Neural Entity Linking	

<https://paperswithcode.com/task/entity-linking>

# OntoNotes

---

- Contains hand-annotated Chinese, Arabic and English coreference datasets
  - From: newswire, magazine articles, broadcast news, broadcast conversations, web data and
  - it does not label singletons (single references)
- Many other datasets reuse from OntoNotes

Reference: <https://catalog.ldc.upenn.edu/LDC2013T19>

Entity Disambiguation is the task of linking mentions of ambiguous entities to their referent entities in a knowledge base such as Wikipedia.

Source: Leveraging Deep Neural Networks and Knowledge Graphs for Entity Disambiguation

# Entity Disambiguation

- State of art and datasets
- Datasets
  - AIDA-CoNLL
  - TAC 2010

## Benchmarks

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE
	AIDA-CoNLL	🏆 confidence-order	Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities		
	MSNBC	🏆 confidence-order	Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities		
	AQUAINT	🏆 MEP + pseudo entities	Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities		
	ACE2004	🏆 confidence-order	Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities		
	WNED-WIKI	🏆 confidence-order	Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities		
	WNED-CWEB	🏆 confidence-order	Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities		
	TAC2010	🏆 NTEE	Learning Distributed Representations of Texts and Entities from Knowledge Base		

<https://paperswithcode.com/task/entity-disambiguation>

# Entity Disambiguation in Practice

## Graph Learning:

1. Knowledge Graph Creation
2. Vertex Embedding Creation

## NLP Execution:

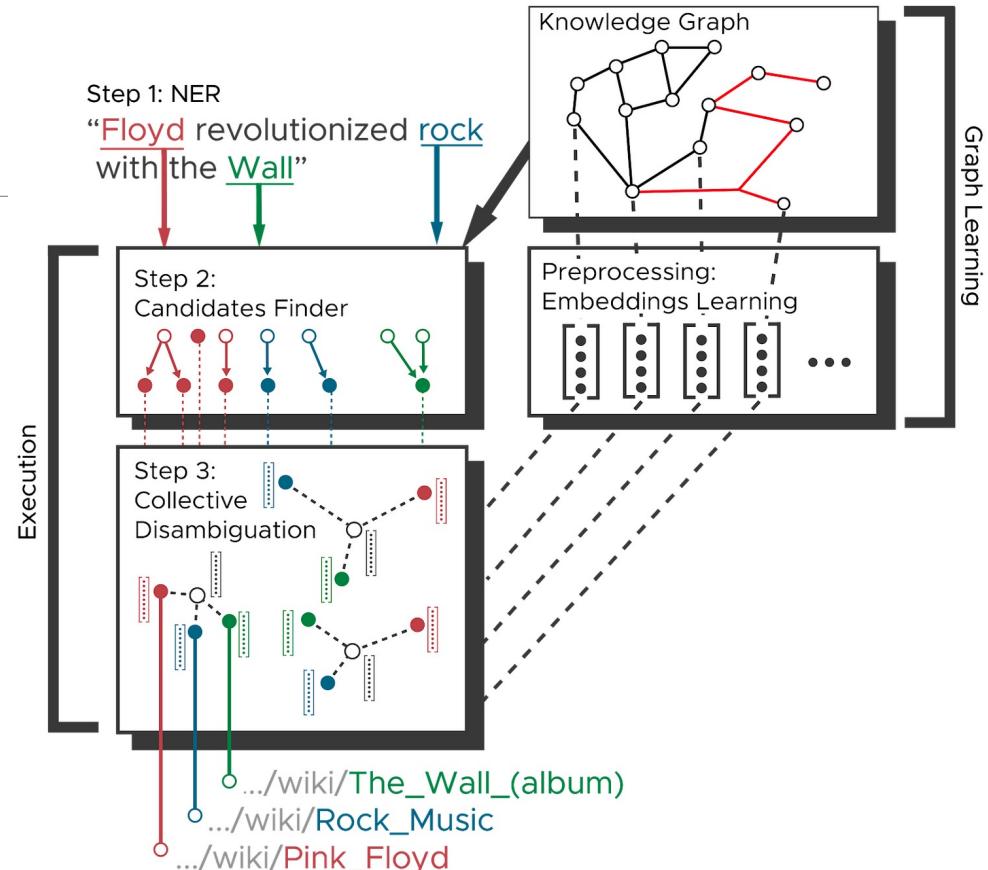
1. Named Entity Recognition
2. Candidate Finder
3. Collective Disambiguation

$$\text{Best Tuple} = \operatorname{argmax}_T \left( \sum_{t_i \in T} \text{Local}(t_i) + \text{Global}(T) \right)$$

Sum of string similarities      Sum of embedding cosine similarities w.r.t. tuple mean

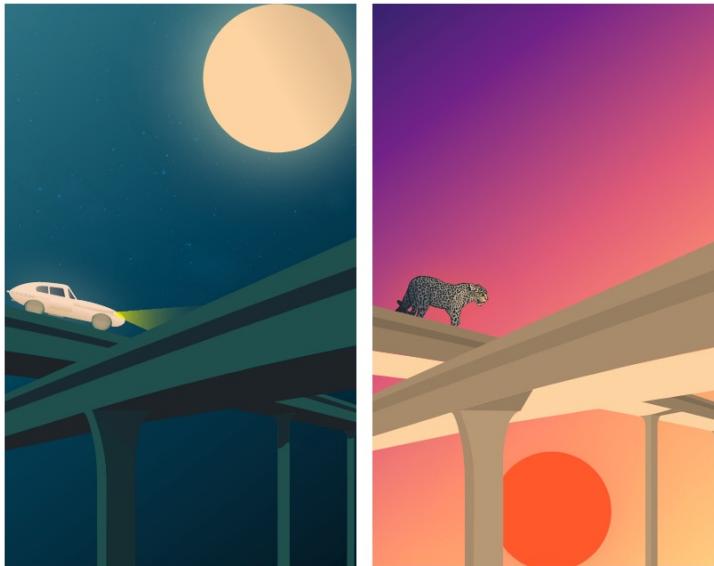
$$\hat{\mathbf{e}}(T) = \frac{1}{|T|} \sum_{t_i \in T} \mathbf{e}(t_i)$$

$$\text{Global}(T) = \sum_{t_i \in T} \frac{\langle \mathbf{e}(t_i), \hat{\mathbf{e}}(T) \rangle}{\|\mathbf{e}(t_i)\|_2 \|\hat{\mathbf{e}}(T)\|_2}$$



Source: <https://blogs.oracle.com/datascience/using-graph-embeddings-for-fast-named-entity-disambiguation-v2>

# Discovering Types and Entity Disambiguation



The man saw a Jaguar speed on the highway.

Jaguar Cars 🚗 0.60

jaguar 🐾 0.29

SEPECAT Jaguar ➔ 0.02

WITHOUT TYPES WITH TYPES

The prey saw the jaguar cross the jungle.

Jaguar Cars 🚗 0.03

jaguar 🐾 0.89

SEPECAT Jaguar ➔ 0.07

WITHOUT TYPES WITH TYPES

Source: <https://openai.com/blog/discovering-types-for-entity-disambiguation/>

# Discovering Types and Disambiguation: Main Steps

---

- Extract every Wikipedia-internal link to determine, for each word, the set of conceivable entities it can refer to. Hence, jaguar animal and car. (Hence, word senses)
- Walk the Wikipedia category tree to determine, for each entity, the set of categories it belongs to
- Pick a list of ~100 categories to be the “type” system.
- Using every Wikipedia-internal link and its surrounding context, produce training data mapping a word plus context to the ~100-dimensional binary representation of the corresponding entity, and train a neural network to predict this mapping.
- At test time, given a word and surrounding context, the neural network’s output can be interpreted as the probability that the word belongs to each category.

**Source:** <https://openai.com/blog/discovering-types-for-entity-disambiguation/>

# Discussion: Named Entity Disambiguation and Word Sense Disambiguation

---

- NED: linking a mention to an actual instance in the given knowledge-base
- WSD: disambiguate words with respect to senses in a directory
- Setting
  - WSD: given word, search in a complete sense directory
  - NED: given candidate, generate all possible candidates
- Synergy:
  - WSD can help in NED, by helping all contexts for an entity

<https://nlp.stanford.edu/pubs/chang2016entity.pdf>

Source: <https://openai.com/blog/discovering-types-for-entity-disambiguation/>

# Case Study: Information Extraction and Linking for Business Documents

---

## References:

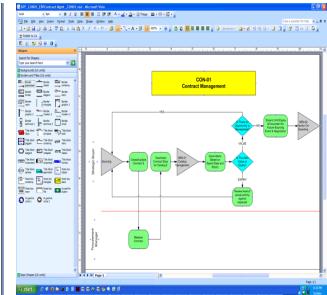
- Elad Fein, Natalia Razinkov, Shlomit Shachor, Pietro Mazzoleni, Sweefen Goh, Richard Goodwin, Manisha Bhandar, Shyh-Kwei Chen, Juhnyoung Lee, Vibha Singhal Sinha, Senthil Mani, Debdoot Mukherjee, Biplav Srivastava, and Pankaj Dhoolia. 2011. Using MATCON to generate CASE tools that guide deployment of pre-packaged applications. In Proceedings of the 33rd International Conference on Software Engineering (ICSE '11). ACM, New York, NY, USA, 1016-1018. DOI:<https://doi.org/10.1145/1985793.1985981>
- Biplav Srivastava, Debdoot Mukherjee, Rema Ananthanarayanan and Vibha Sinha, Model Extraction to Model-based Reuse of Enterprise Documents, In ACM International Conference on Management of Data (COMAD 2010), Nagpur, India, 8-10 December 2010. [Area:Content Harvesting-Model extraction]
- Debdoot Mukherjee ; Senthil Mani ; Vibha Singhal Sinha ; Rema Ananthanarayanan ; Biplav Srivastava, Pankaj Dhoolia ; Prahlad Chowdhury, "AHA: Asset Harvester Assistant," 2010 IEEE International Conference on Services Computing, Miami, FL, 2010, pp. 425-432. [Area:Content Harvesting-Services]
- Biplav Srivastava, Yuan-chi Chang, Business Insight from Collection of Unstructured Formatted Documents with IBM Content Harvester, In ACM International Conference on Management of Data (COMAD 2009), Mysore, India, 9-12 December 2009. [Area:BusinessProcess-DocHarvesting]
- Biplav Srivastava, Debdoot Mukherjee, Organizing Documented Processes, In IEEE International Conference on Services Computing (SCC 2009), Bangalore, India, September 21-25, 2009.[Area:BusinessProcess-TextFlowAnalyses]

# Key Information Elements

## Process Specific Artifacts

- Scenario
- Process
- Process Step
- Inputs, Outputs

ERP Project		
Process Definition Document		
Customers		
Actions	By Who	Managerial System
23.1 Document whatever is there in the Americas	Manager	Manual
23.2 Send preexisting request emails to relevant parties (eg. HR Service Center, Local HR, Local Helpdesk, Local IS, Local R&D and generate work orders. Get to 2.1.4 and 2.1.5)	HR Service Center, Local HR	Manual
23.3 Approve and inform relevant providers (eg. relevant HR Service Center, Local HR, Local Helpdesk, Local R&D and phone and notify manage accordingly. Get to 2.4 (Complete New Hire Orientation process))	Other	Manual
23.4 Approve and inform relevant providers (eg. relevant HR Service Center, Local HR, Local Helpdesk, Local R&D and phone and notify manage accordingly. Get to 2.4 (Complete New Hire Orientation process))	Other	Manual



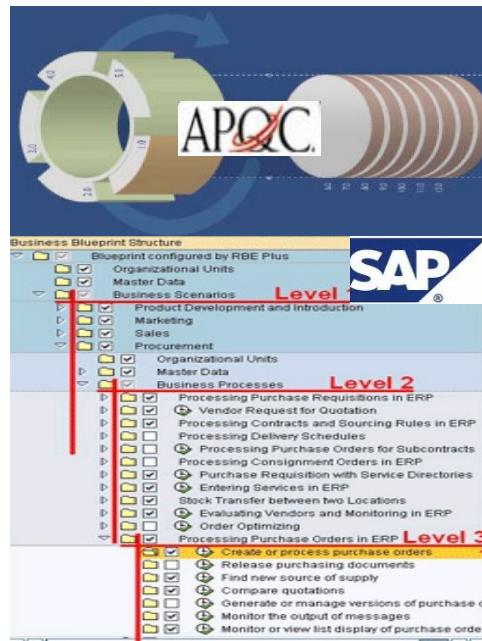
## Non-Process Business Artifacts

- Requirement
- Use-case
- Gap
- KPI

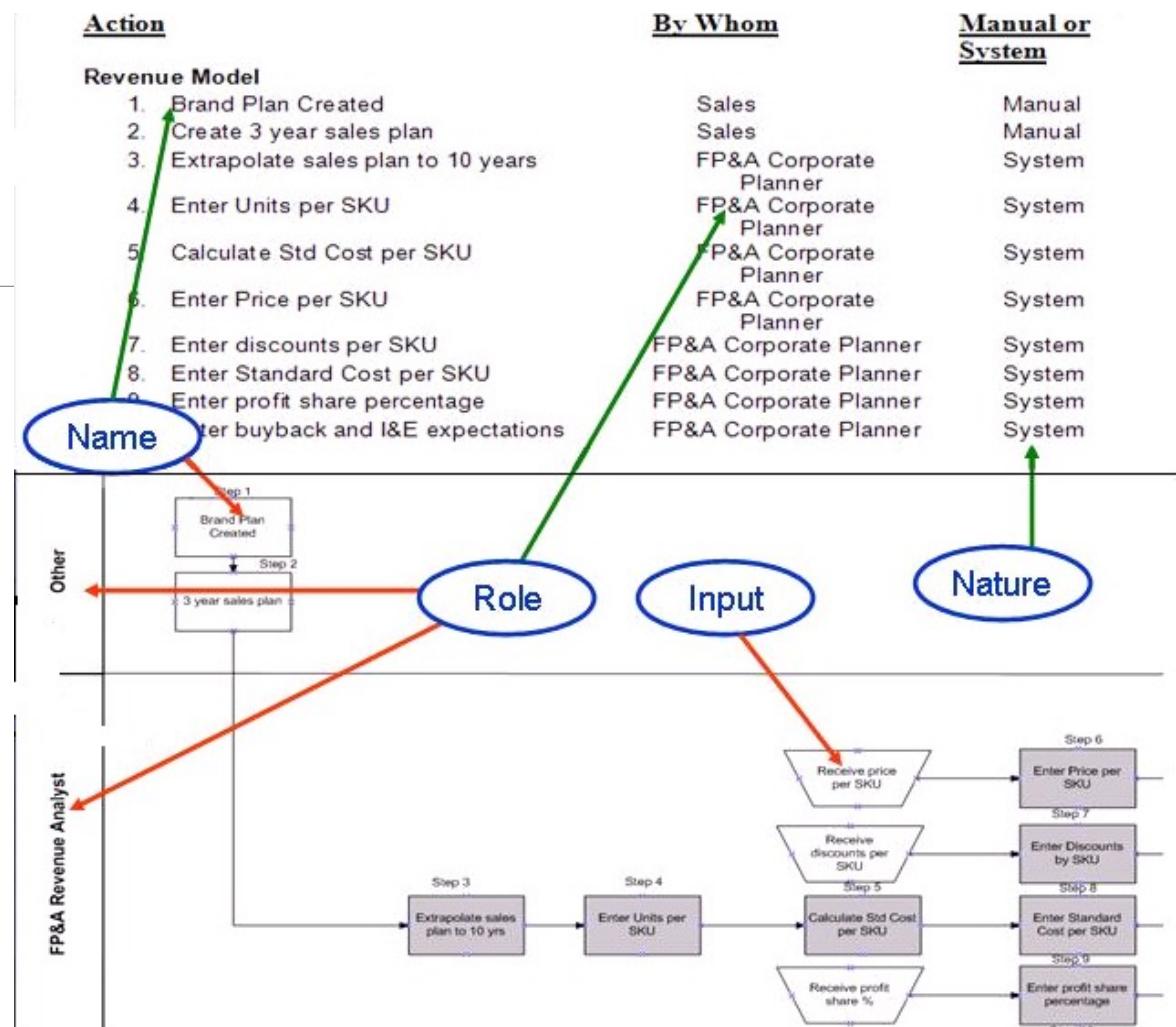


## Business Process Hierarchies

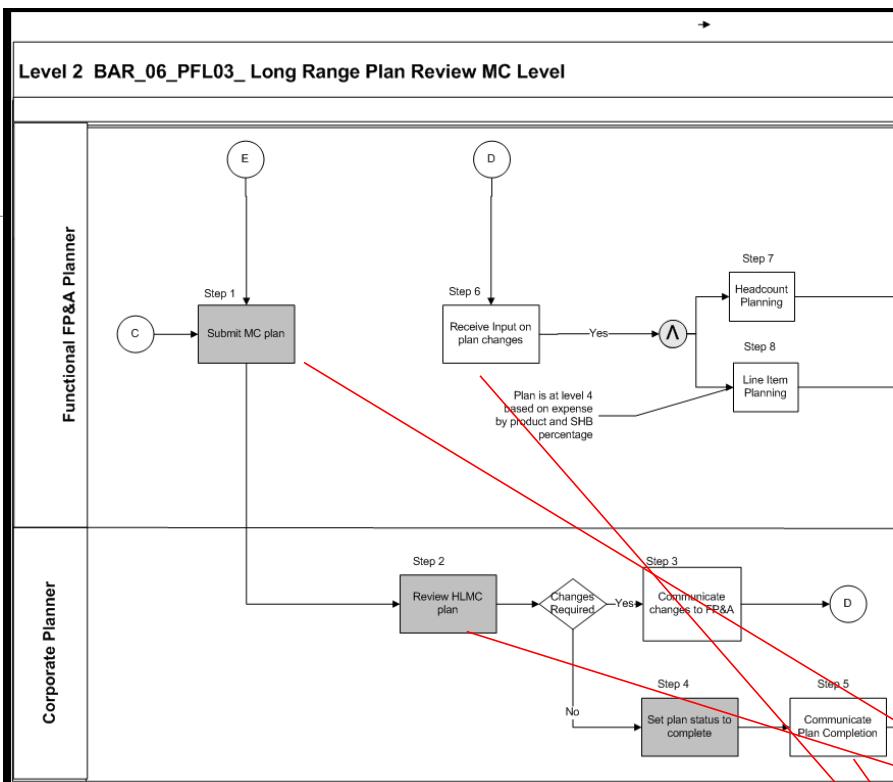
- Industry Specific
- Cross Industry



## Process Information in Text and Visual Formats



# Example



Benefits in flow:

- Control flow is detailed
- Intuitive

**\* Problems in flow:**

- Names in flow do not match text (Functional FP&A Planner v/s (FP&A Planner))
- Limited information. E.g., whether an activity is system or manual?

Text has the details

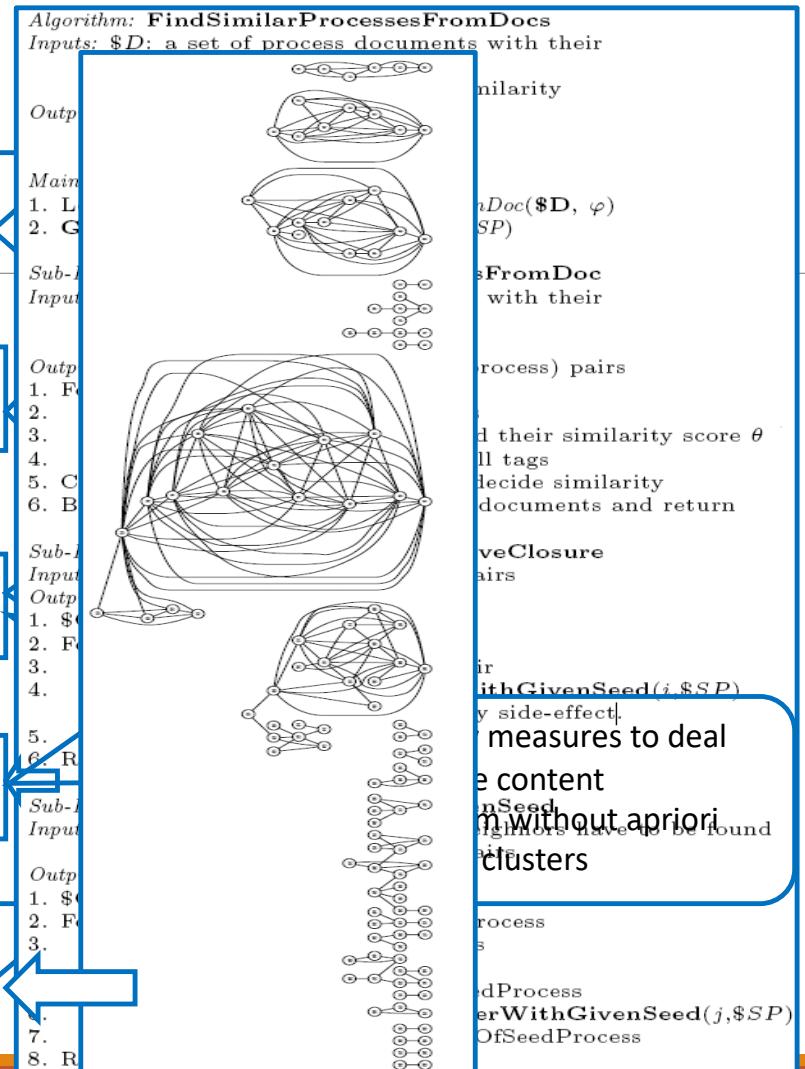
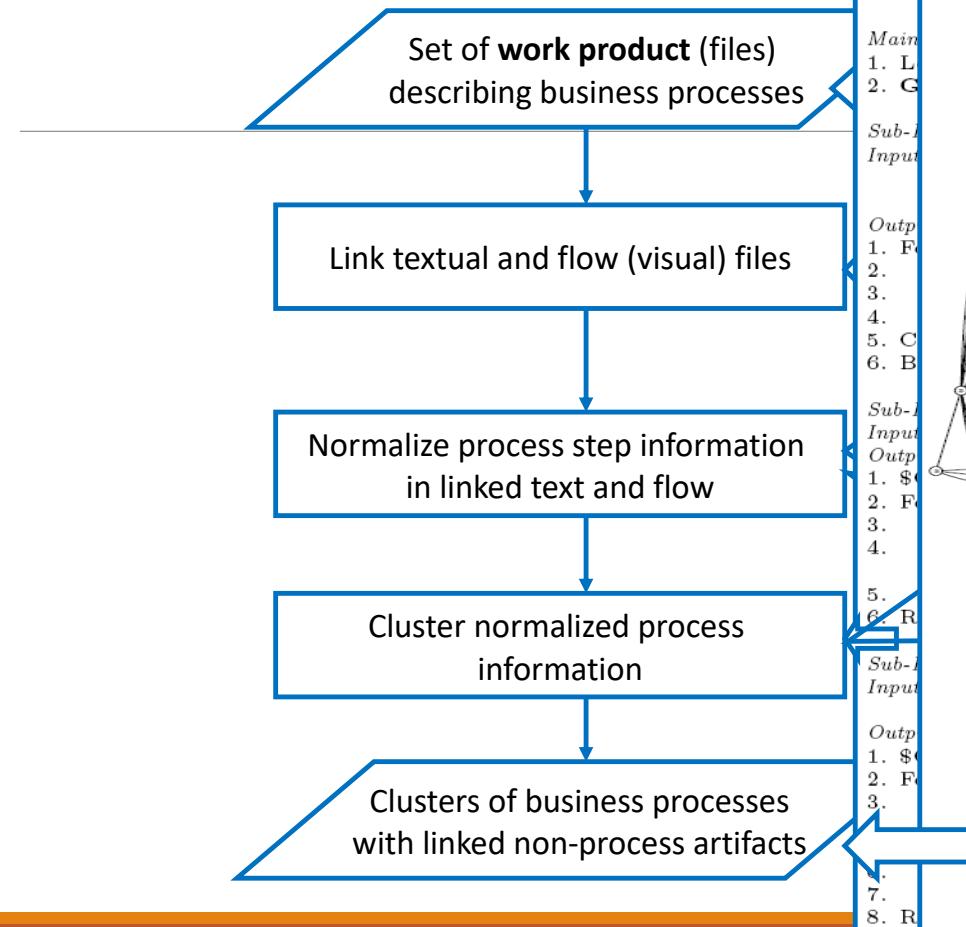
Mozilla Firefox

File Edit View History

Steps: BusinessProcess.procStep

Action	Revenue Model	
Brand Plan Created	Sales	Manual
Create 3 year sales plan	Sales	Manual
Extrapolate sales plan to 10 years	FP&A Corporate Planner	System
Enter Units per SKU	FP&A Corporate Planner	System
Calculate Std Cost per SKU	FP&A Corporate Planner	System
Enter Price per SKU	FP&A Corporate Planner	System
Enter discounts per SKU	FP&A Corporate Planner	System
Enter Standard Cost per SKU	FP&A Corporate Planner	System
Enter profit share percentage	FP&A Corporate Planner	System
Enter buy/back and I&E expectations	FP&A Corporate Planner	System
Prepare Baseline		
Copy latest estimate into plan version	FP&A Corporate Planner	System
Copy annual plan	FP&A Corporate Planner	System
Input Long Range plan	FP&A Corporate Planner	System
Calculate revenue and cost of goods sold	FP&A Corporate Planner	System
Input Capex plan	FP&A Corporate Planner	System
Input Commitments	FP&A Corporate Planner	System
Review LRP	FP&A Corporate Planner	System
Adjust LRP	FP&A Corporate Planner	System
Release for OPEX planning	FP&A Corporate Planner	Manual
Plan Review HLMC level		
Submit HLMC plan	FP&A Planner	System
Review HLMC plan	Corporate Planner	System
Communicate Changes to FP&A	Corporate Planner	Manual
Set plan status to complete	Corporate Planner	System
Communicate plan completion	Corporate Planner	System
Receive input on plan changes	FP&A Planner	Manual
Input Headcount planning	FP&A Planner	Manual
Input Line item planning	FP&A Planner	Manual

# Linking Steps in Process Organization

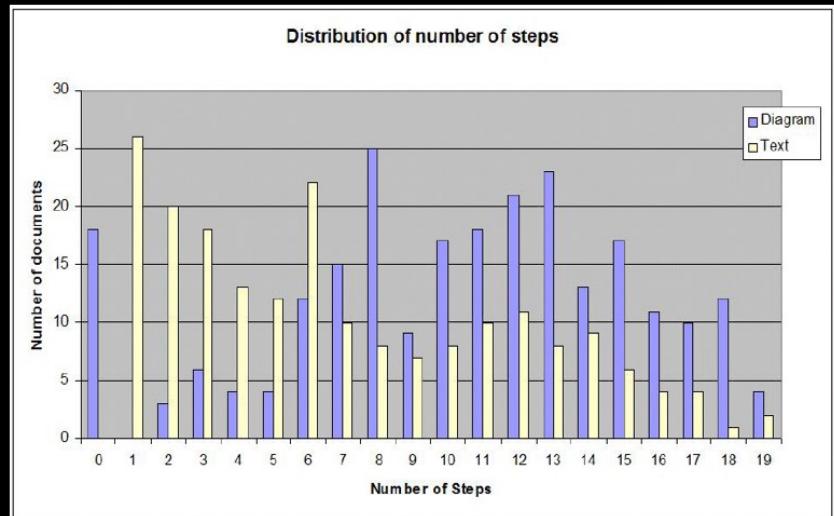


# Empirical Evaluation — Results

- Input
  - 240 Process Definition Documents
  - 315 Process Flow Diagrams

- Linking

Similarity Measure	Pair-wise Matches	# PDDs	Precision (%)
Jaro	126	30	48
Exact	11	11	100



- Normalization

Similarity Measure	% Match (Name)	% Match (Name + Role)
Jaro	37	8
Exact	45.5	13

# Lecture 17: Concluding Comments

---

- We looked at
  - Entity extraction
  - Extraction linking
- Considered the use-case of documents in packaged application customization projects

# Concluding Segment

---

# About Next Lecture – Lecture 18

---

# Lecture 18 Outline

---

- Event extraction
- Spatio-temporal analysis