AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 18: Entity Extraction, Linking

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

22ND OCTOBER, 2024

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 18

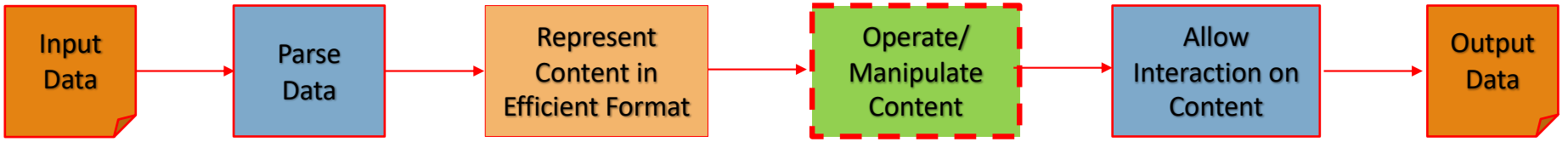
- Opening Segment
  - Announcements
- Main Lecture

Main Section
- Entity Extraction
- Entity Linking

- Concluding Segment
  - About Next Lecture – Lecture 19

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

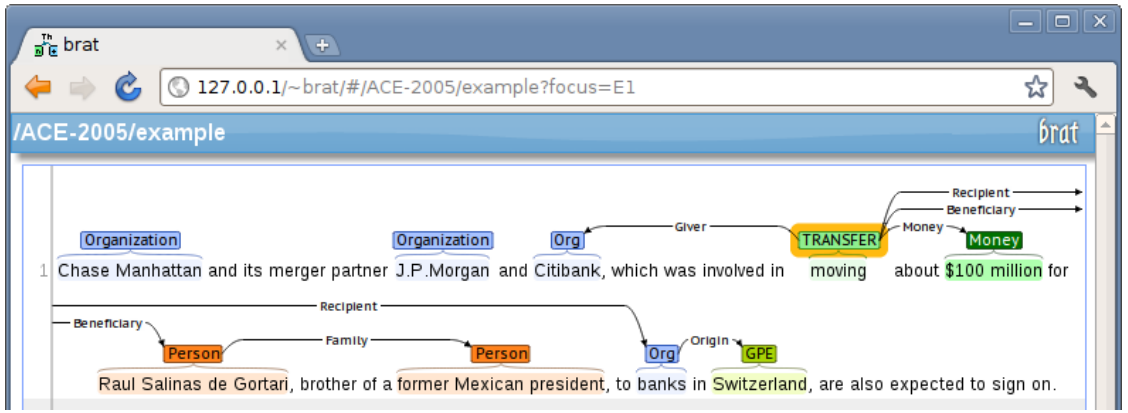
The **Nobel Peace Prize** is one of the five **Nobel Prizes** established by the **will** of Swedish industrialist, inventor, and armaments manufacturer **Alfred Nobel**, along with the prizes in Chemistry, Physics, Physiology or Medicine, and Literature

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

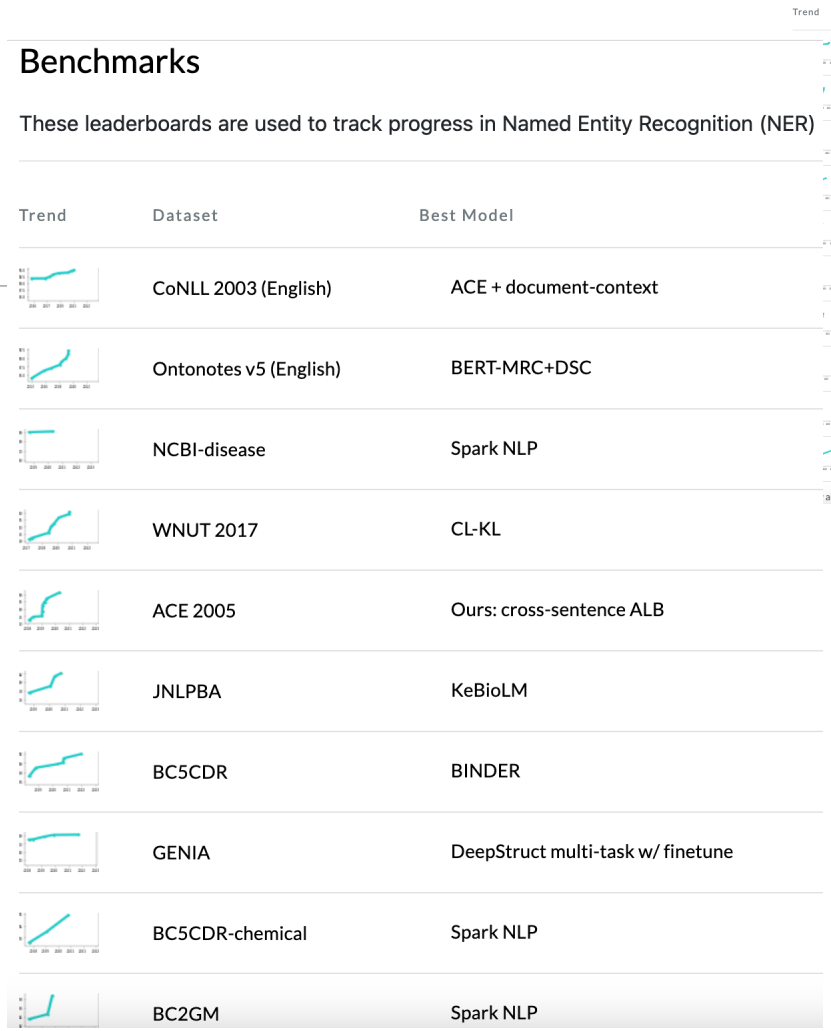
Reference: https://lionbridge.ai/articles/the-essential-guide-to-entity-extraction/

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 – Oct 2022

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

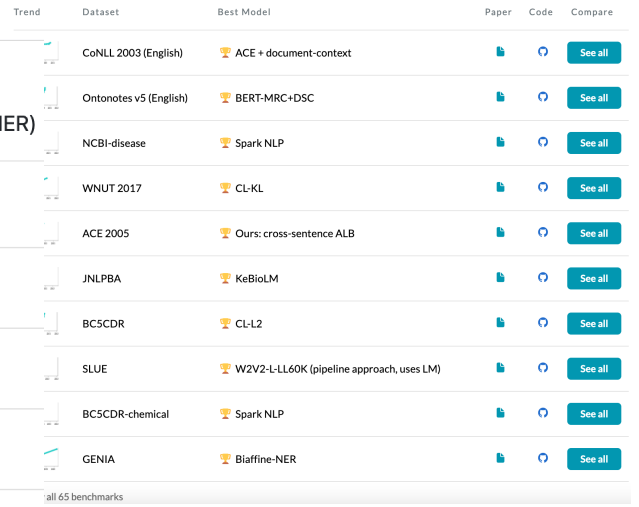


Benchmarks

These leaderboards are used to track progress in Named Entity Recognition (NER)

Trend	Dataset	Best Model
	CoNLL 2003 (English)	ACE + document-context
	Ontonotes v5 (English)	BERT-MRC+DSC
	NCBI-disease	Spark NLP
	WNUT 2017	CL-KL
	ACE 2005	Ours: cross-sentence ALB
	JNLPBA	KeBioLM
	BC5CDR	BINDER
	GENIA	DeepStruct multi-task w/ finetune
	BC5CDR-chemical	Spark NLP
	BC2GM	Spark NLP

Oct 2024



Trend	Dataset	Best Model	Paper	Code	Compare
	CoNLL 2003 (English)	ACE + document-context			See all
	Ontonotes v5 (English)	BERT-MRC+DSC			See all
	NCBI-disease	Spark NLP			See all
	WNUT 2017	CL-KL			See all
	ACE 2005	Ours: cross-sentence ALB			See all
	JNLPBA	KeBioLM			See all
	BC5CDR	CL-L2			See all
	SLUE	W2V2-L-LL60K (pipeline approach, uses LM)			See all
	BC5CDR-chemical	Spark NLP			See all
	GENIA	Biaffine-NER			See all

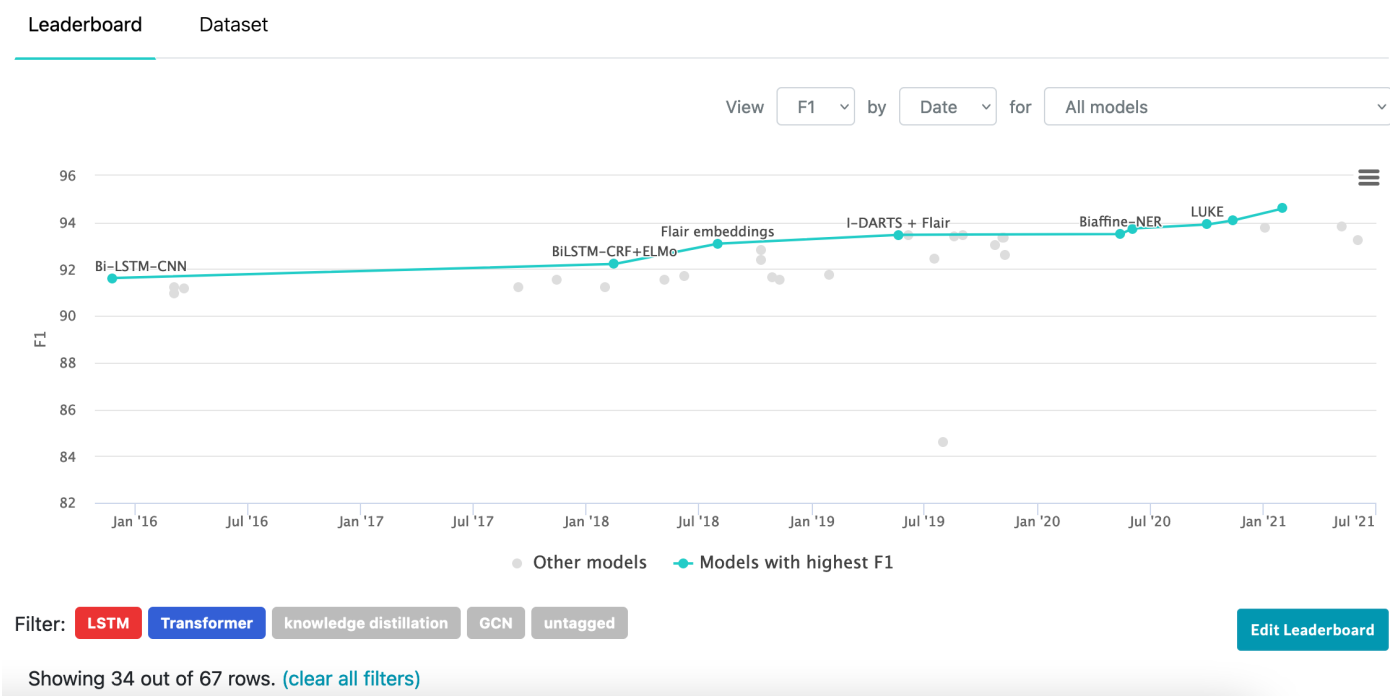
all 65 benchmarks

Oct 2022

Benchmark on a Dataset – Oct 2022

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

Named Entity Recognition on CoNLL 2003 (English)



Benchmarks – Oct 2020

- 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			See all
	Ontonotes v5 (English)	BERT-MRC+DSC	Dice Loss for Data-imbalanced NLP Tasks			See all
	ACE 2005	BERT-MRC	A Unified MRC Framework for Named Entity Recognition			See all
	GENIA	BERT-MRC	A Unified MRC Framework for Named Entity Recognition			See all
	CoNLL++	CrossWeigh + Pooled Flair	CrossWeigh: Training Named Entity Tagger from Imperfect Annotations			See all
	Long-tail emerging entities	Flair embeddings	Contextual String Embeddings for Sequence Labeling			See all
	BC5CDR	NER+PA+RL (PubMed)	Reinforcement-based denoising of distantly supervised NER with partial annotation			See all
	JNLPBA	BioBERT	BioBERT: a pre-trained biomedical language representation model for biomedical text mining			See all
	SciERC	SpERT	Span-based Joint Entity and Relation Extraction with Transformer Pre-training			See all

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/

Coding Resource

- **Notebook**: https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l17-eventextr/SimpleEntitySearch.ipynb

BioMedical Entity Extaction

- 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:

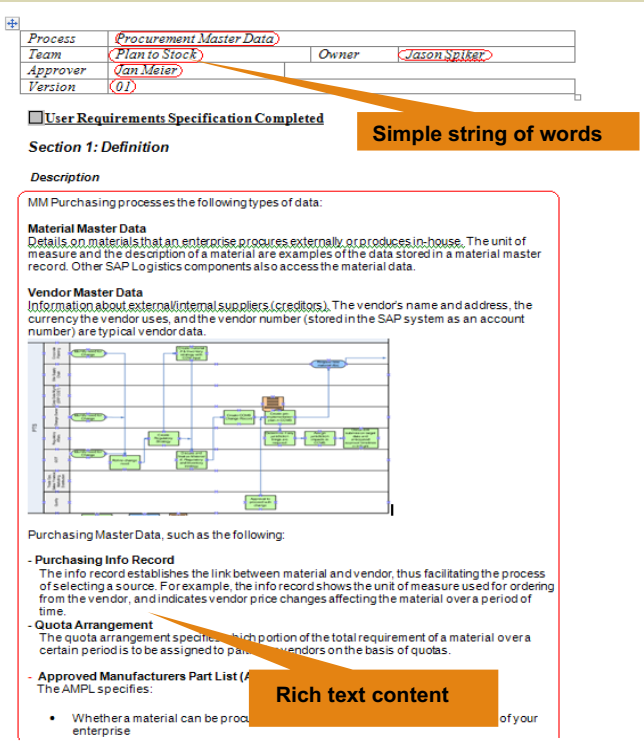
- Project page: https://sites.google.com/site/biplavsrivastava/research-1/packaged-software-services
- Elad Fein, Natalia Razinkov, Shlomit Shachor, Pietro Mazzoleni, Sweefen Goh, Richard Goodwin, Manisha Bhandar, Shyh-Kwei Chen, Juhnyoung Lee, Vibha Singhal Sinha, Senthil Mani, Debdoot Mukherjee, Biplav 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 Harveting-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	Procurement Master Data
Team	Plan to Stock
Owner	Jason Spiker
Approver	Jan Meier
Version	01

☐ **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 (A**
  The AMPL specifies:
  - Whether a material can be procured for your enterprise

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 plant
- 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
- PD
- Planner
- Planner/Buyer
- Supplier
- Finance

Outputs

Master Data Record Created for the views and

Customers

- Planner/Buyers, Planners, Manufacturing, M

Steps

Detail description of each of the boxes in the pr

Action	nom	Manual or System
1. Maintain Material Master	Corporate Data Management, SQM,	System
2. Maintain Vendor Master	Planner/Buyer Finance, SQM,	System

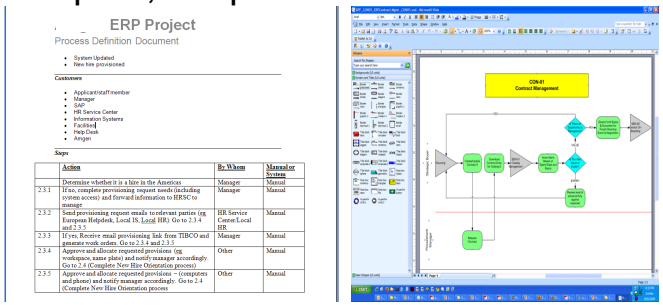
Simple List extraction unit: Each record in this list maps to a new instance of a particular entity in the model.

Complex entity extraction unit. Each record in the tables maps to anew instance of a particular entity in the model. Each column in turn maps to a particular attribute in the entity

Key Information Elements

Process Specific Artifacts

- Scenario
- Process
- Process Step
- Inputs, Outputs


ERP Project
Process Definition Document

- System Updated
- New hire provisioned

Customers

- Applicant/staff member
- Manager
- SAP
- HR Service Center
- Information Systems
- Facilities
- Help Desk
- Amgen

Steps

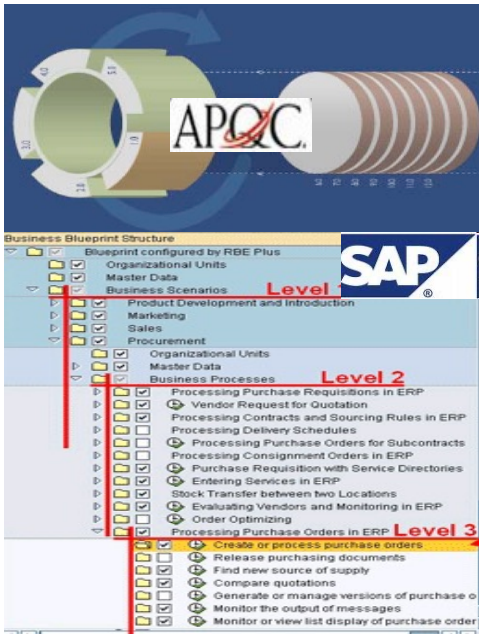
	Action	By Whom	Manual or System
	Determine whether it is a hire in the Americas	Manager	Manual
2.3.1	If no, complete provisioning request needs (including system access) and forward information to HRSC to manage	Manager	Manual
2.3.2	Send provisioning request emails to relevant parties (eg European Helpdesk, Local IS, Local HR). Go to 2.3.4 and 2.3.5	HR Service Center/Local HR	Manual
2.3.3	If yes, Receive email provisioning link from TIBCO and generate work orders. Go to 2.3.4 and 2.3.5	Manager	Manual
2.3.4	Approve and allocate requested provisions (eg workspace, name plate) and notify manager accordingly. Go to 2.4 (Complete New Hire Orientation process)	Other	Manual
2.3.5	Approve and allocate requested provisions – (computers and phone) and notify manager accordingly. Go 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



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

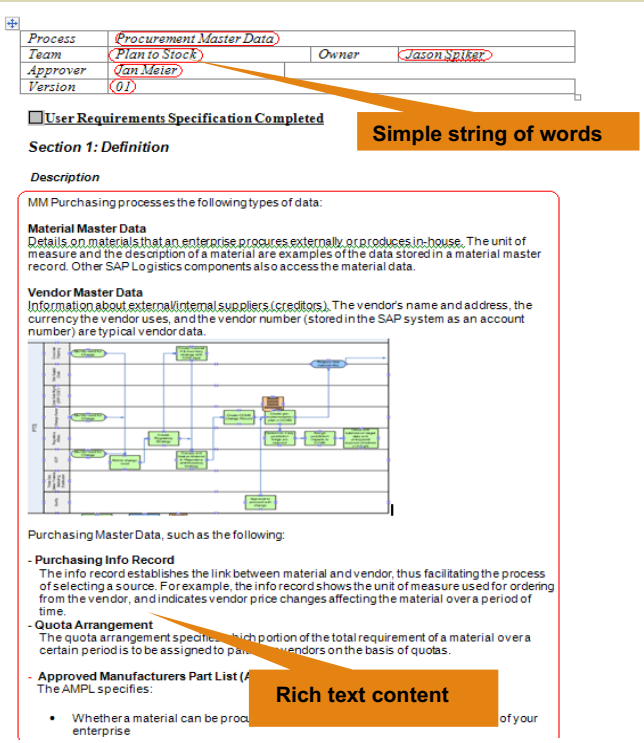
- More than a decade of literature exists in information extraction from Web data
- *Wrappers* are programs to extract information from web pages
- Broadly 3 categories of information extractors:
  - Rule-based: Extraction patterns are specified using some specialized notation
  - Supervised: Wrappers 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	Procurement Master Data
Team	Plan to Stock
Owner	Jason Spiker
Approver	Jan Meier
Version	01

☐ **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 (A**
  The AMPL specifies:
  - Whether a material can be procured for your enterprise

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 plant
- 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
- PD
- Planner
- Planner/Buyer
- Supplier
- Finance

Outputs

Master Data Record Created for the views and

Customers

- Planner/Buyers, Planners, Manufacturing, M

Steps

Detail description of each of the boxes in the pr

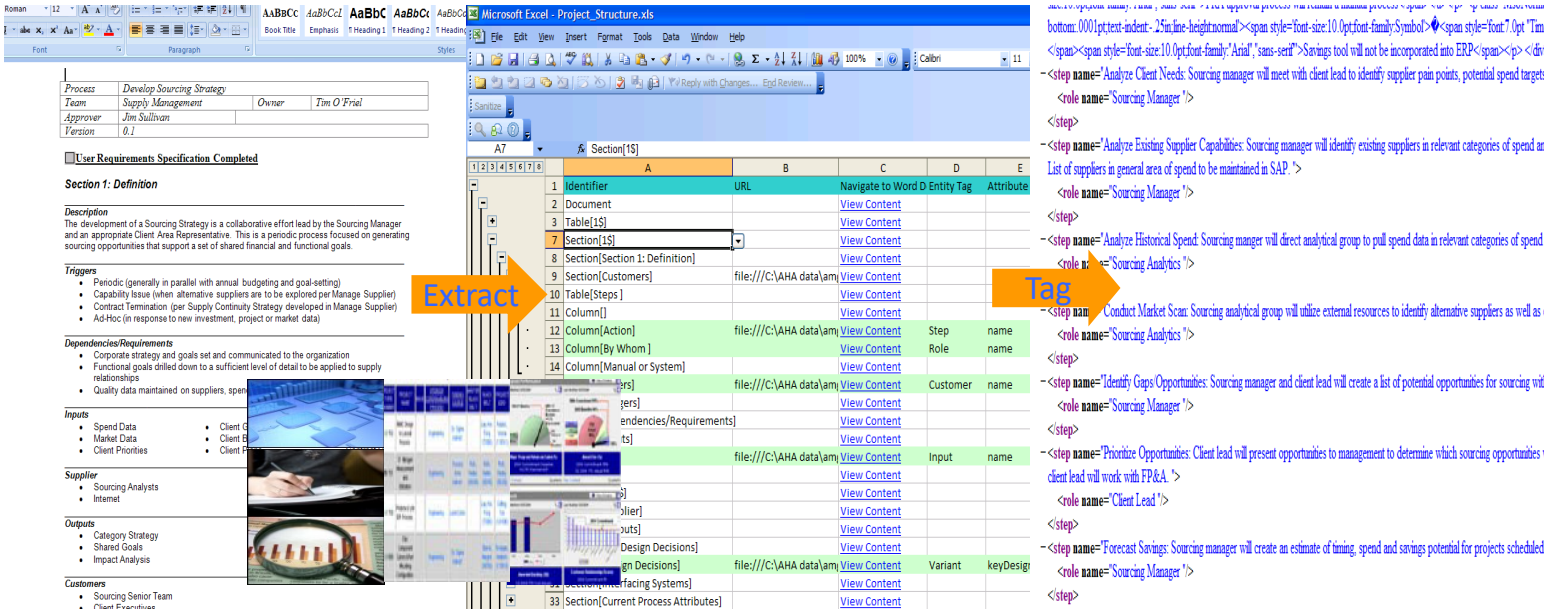
Action	nom	Manual or System
1. Maintain Material Master	Corporate Data Management, SQM,	System
2. Maintain Vendor Master	Planner/Buyer Finance, SQM,	System

Simple List extraction unit: Each record in this list maps to a new instance of a particular entity in the model.

Complex entity extraction unit. Each record in the tables maps to anew instance of a particular entity in the model. Each column in turn maps to a particular attribute 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



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/anhaidgroup/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

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$_i$ and told him$_i$ about a new patient [pro-stereotypical]

The secretary called the physician$_i$ and told her$_i$ 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
Walmart	Debug a system in production that matches products	Improved recall by 34%, reduced precision by 0.65%	Yes	1 student, 1 employee	Funding
Recruit holdings	Matching names of stores, companies, properties	Reported 98.9% accuracy on matching 10K store names	Yes	Multiple employees	Press release
Johnson controls	Matching suppliers	Precision and recall in 96–100%	Unknown	1 student	Funding
Marshfield clinic	Matching drugs	99.2% precision and 95.3% recall	Yes	1 student, 1 employee	Paper
Economics (UW)	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
Land use (UW)	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
Biomedicine (UW)	Matching ontology terms	metasra.biostat.wisc.edu	Yes	1 student	Paper published
Limnology (UW)	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/anhaidgroup/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 – Gazeteer

Wikipedia Example: White House

https://en.wikipedia.org/wiki/White_House

Contents [hide]

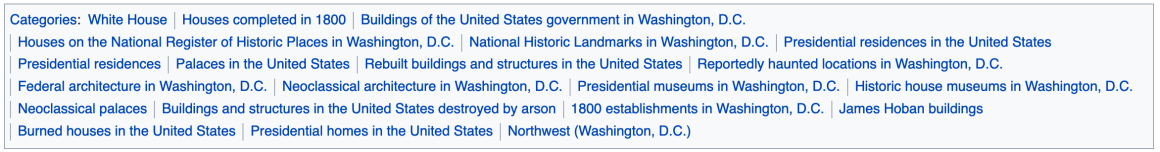
1 Early history
- 1.1 1789–1800
- 1.2 Architectural competition
- 1.3 Design influences
- 1.4 Construction
- 1.5 Architectural description
- 1.6 Naming conventions

2 Evolution of the White House
- 2.1 Early use, the 1814 fire, and rebuilding
- 2.2 Overcrowding and building the West Wing
- 2.3 Truman reconstruction
- 2.4 Jacqueline Kennedy restoration

3 The White House since the Kennedy restoration
- 3.1 Layout and amenities
- 3.2 Executive Residence
- 3.3 West Wing
- 3.4 East Wing
- 3.5 Grounds

4 Public access and security
- 4.1 Historical accessibility
- 4.2 Aviation incidents
- 4.3 Closure of Pennsylvania Avenue
- 4.4 Protection

5 See also
6 Notes
7 References
8 Further reading
9 External links



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.)

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

Wikipedia Example: White House

https://en.wikipedia.org/wiki/White_House

Document context:

- Mentions
- Named entities
- Coreferences
- disambiguations

Wikipedia context:

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

Contents [hide]

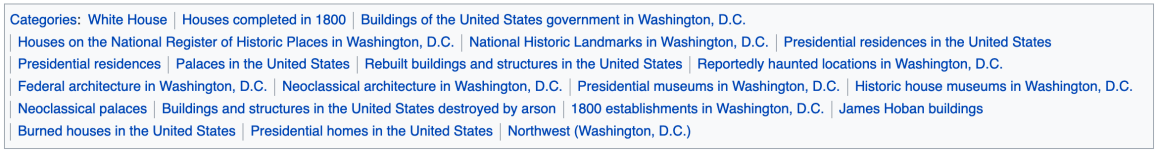
1 Early history
- 1.1 1789–1800
- 1.2 Architectural competition
- 1.3 Design influences
- 1.4 Construction
- 1.5 Architectural description
- 1.6 Naming conventions

2 Evolution of the White House
- 2.1 Early use, the 1814 fire, and rebuilding
- 2.2 Overcrowding and building the West Wing
- 2.3 Truman reconstruction
- 2.4 Jacqueline Kennedy restoration

3 The White House since the Kennedy restoration
- 3.1 Layout and amenities
- 3.2 Executive Residence
- 3.3 West Wing
- 3.4 East Wing
- 3.5 Grounds

4 Public access and security
- 4.1 Historical accessibility
- 4.2 Aviation incidents
- 4.3 Closure of Pennsylvania Avenue
- 4.4 Protection

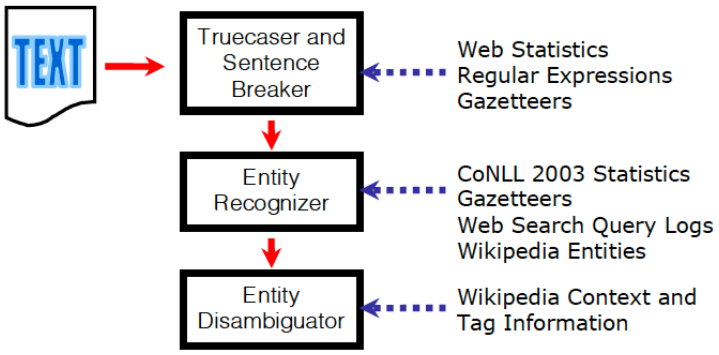
5 See also
6 Notes
7 References
8 Further reading
9 External links



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

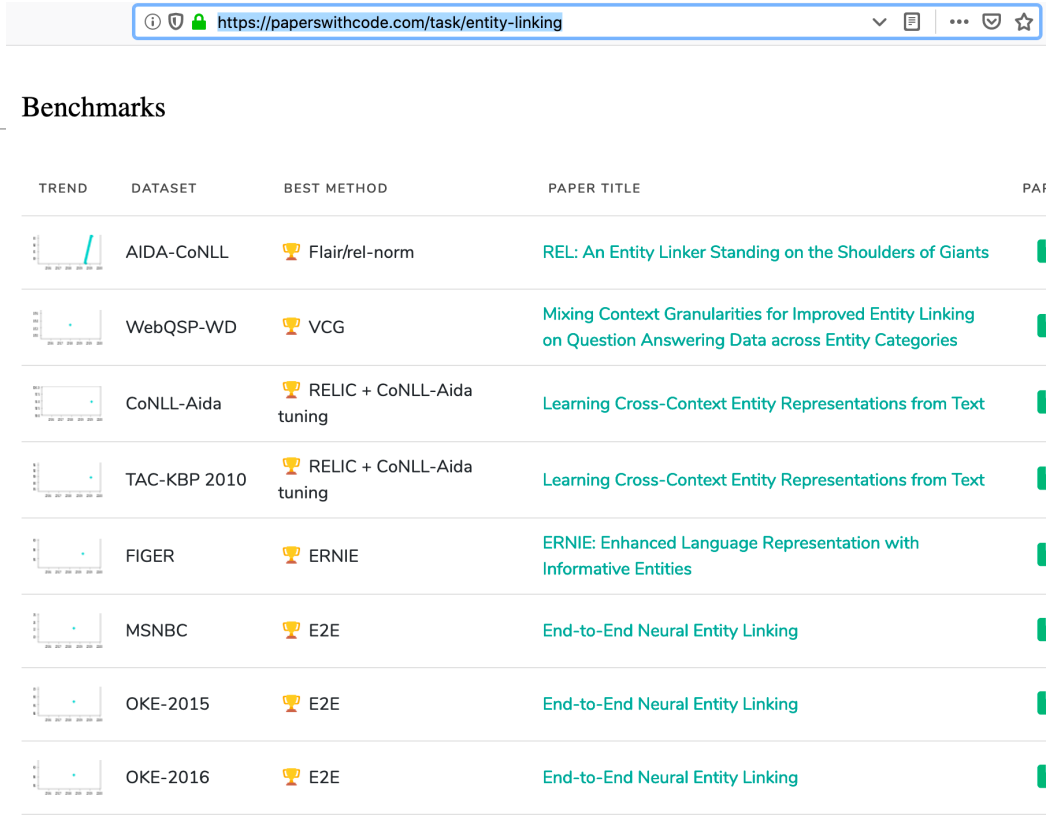


Approach:

- 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

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

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

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

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

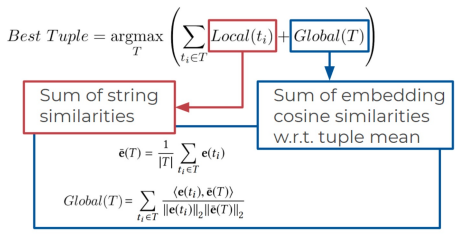
Entity Disambiguation in Practice

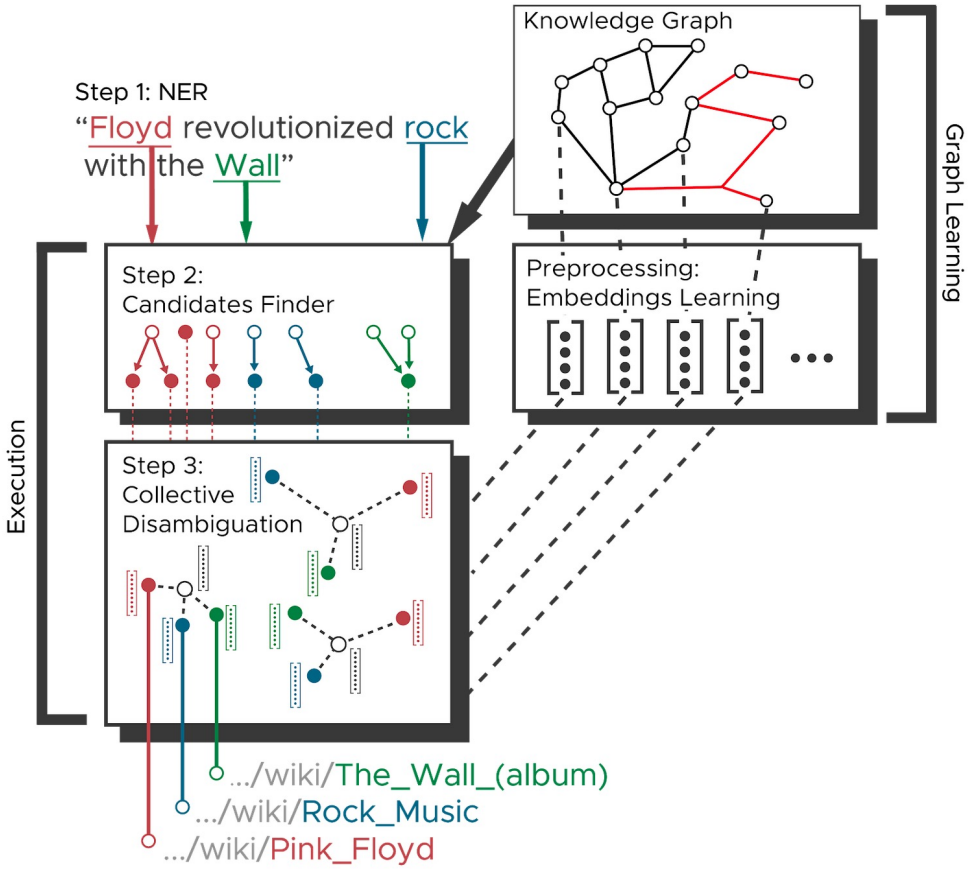
Graph Learning:

1. Knowledge Graph Creation
2. Vertex Embedding Creation

NLP Execution:

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

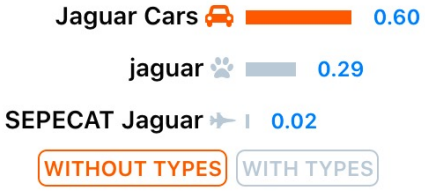




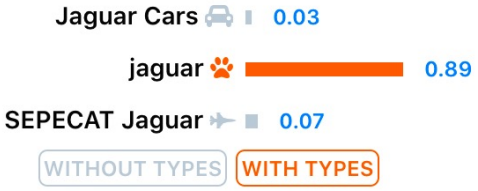
Source: https://blogs.oracle.com/datascience/using-graph-embeddings-for-fast-named-entity-disambiguation-v2

Discovering Types and Entity Disambiguation

The man saw a **Jaguar** speed on the highway.



The prey saw the **jaguar** cross the jungle.



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

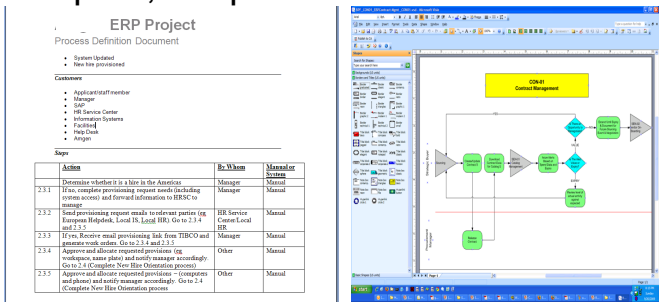
References:

- Project page: https://sites.google.com/site/biplavsrivastava/research-1/packaged-software-services
- 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 Harveting-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

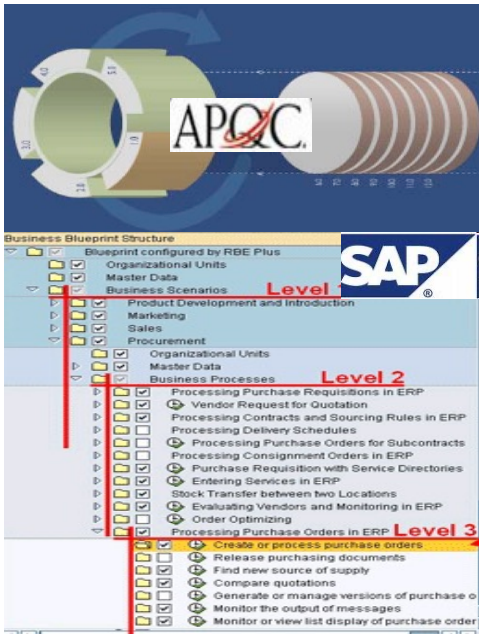


Non-Process Business Artifacts

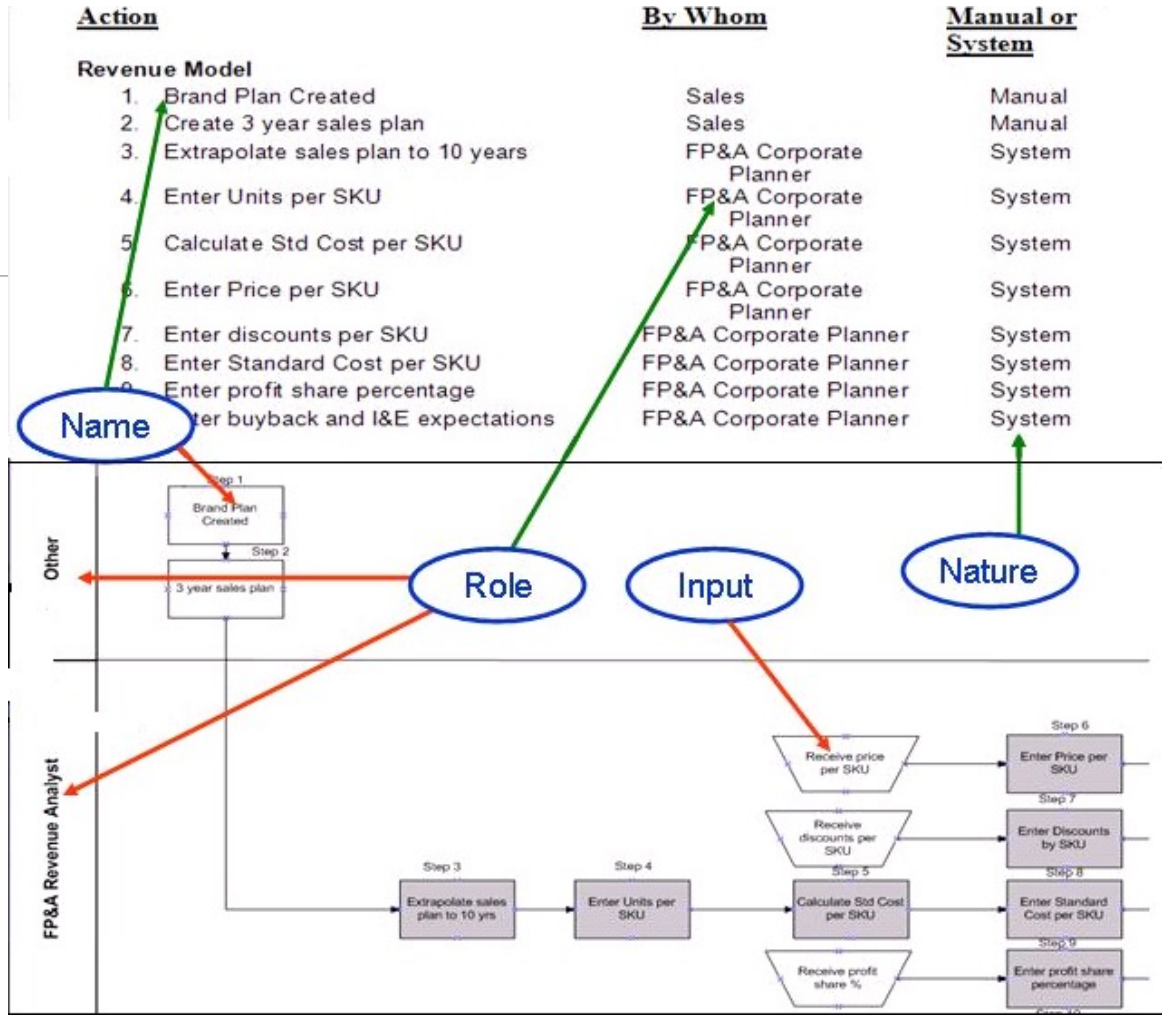
- Requirement
- Use-case
- Gap
- KPI

Business Process Hierarchies

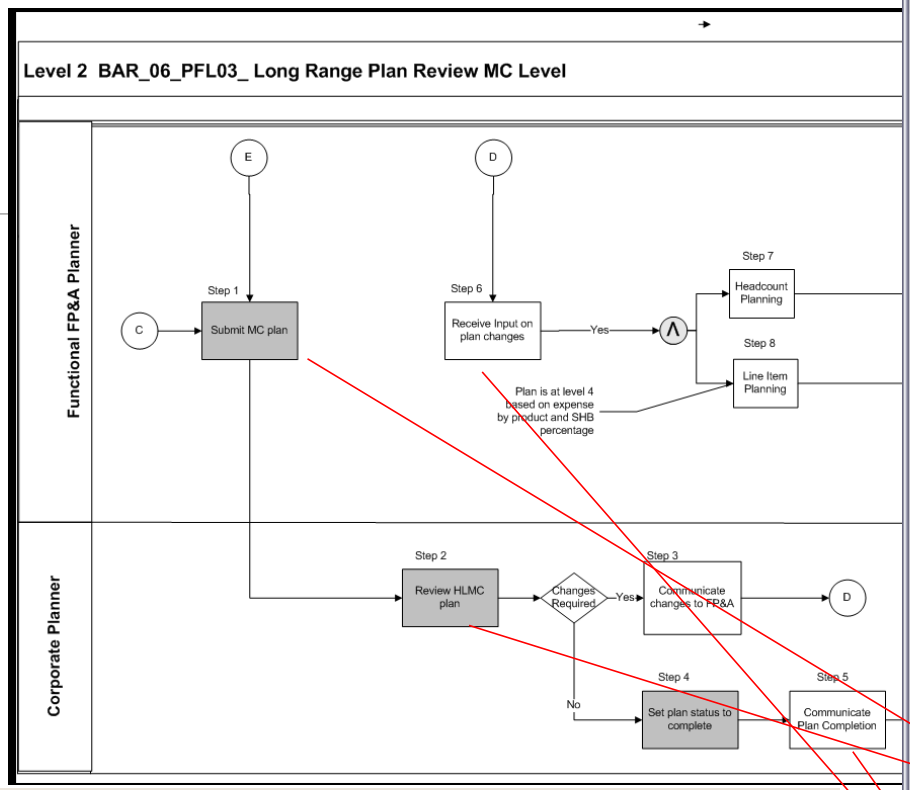
- Industry Specific
- Cross Industry

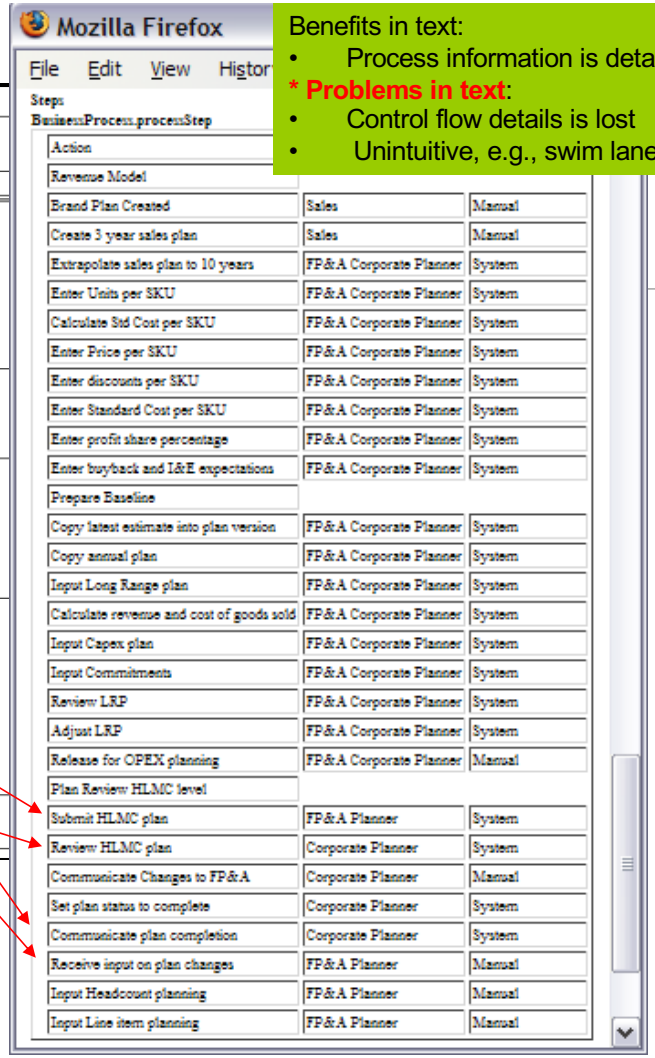


Process Information in Text and Visual Formats



Example





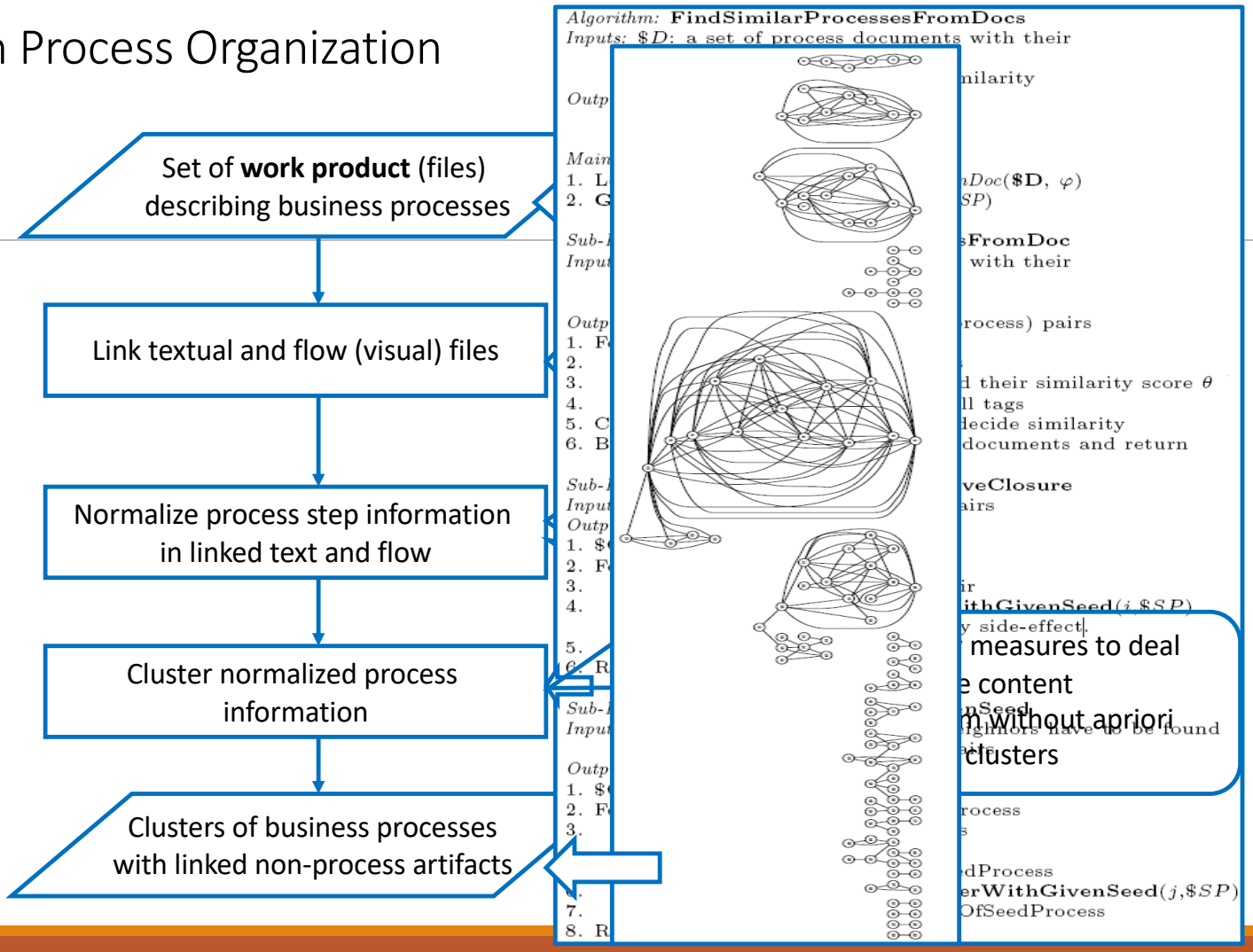
Benefits in text:
- Process information is detailed

*** Problems in text**:
- Control flow details is lost
- Unintuitive, e.g., swim lanes is missing

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

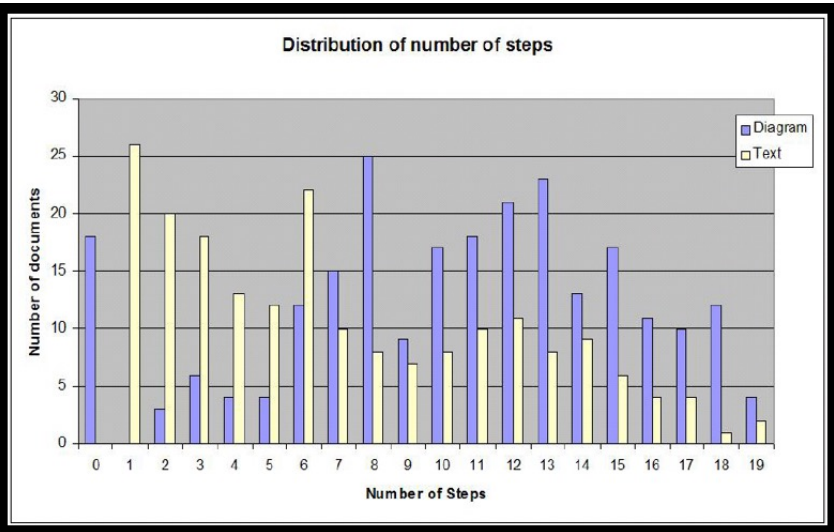
Linking Steps in Process Organization



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 18: Concluding Comments

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

Concluding Segment

Discussion: Course Project

Theme: Analyze quality of official information available for elections in 2024 [in a state]

- Take information available from
  - Official site: State Election Commissions
  - Respected non-profits: League of Women Voters
- Analyze information
  - State-level: Analyze quality of questions, answers, answers-to-questions
  - Comparatively: above along all states (being done by students)
- Benchmark and report
  - Compare analysis with LLM
  - Prepare report

- Process and analyze using NLP
  - Extract entities
  - Assess quality – metrics
    - Content – *Englishness*
    - Content – *Domain* -- election
  - … other NLP tasks
  - Analyze and communicate overall

Major dates for project check

- Sep 10: written – project outline
- Oct 8: in class
- Oct 31: in class // LLM
- Dec 5: in class // Comparative

Obtaining Election Data

Here are a few things to do:
A) **Official data** backed by laws: state election commission
a) Find the state's election commission
b) Find the Q/As they provide. They may be as FAQs or on different web pages.
c) Collect the Q/A programmatically

B) **Secondary data** sources: non-profit
a) Find Q/As from Vote 411 which is supported by the non-profit: LWV.

For reference, for SC,
A) Official - https://scvotes.gov/voters/voter-faq/
B) Secondary - https://www.vote411.org/south-carolina

For extraction, one or more approaches:
- Manually annotating
- BeautifulSoup,
- Tika
- or other open source libraries.

Election Q/A for Your State

- Format in .json; name file as "**xy**_qa.json", where xy is the two-character US state acronym
- Fixed attributes in .json
  - state: xy
  - num_questions: **a**, where a is the number of questions
  - num_answers: **b**, where b is the number of answers
  - contributor: student name
- questions: List of Q/As with attributes for each it:
  - q // question
  - a // answer
  - s // source url from where the information is taken
  - t // time when the information is taken – UTC format
- Store it in your github repo; put in sub-dir like "project/data"
- Instructor will keep it in common place inside course github repo and share.

Election Q/As for Multiple States

- Instructor will keep it common place inside course github repo and share.
- You will be able to access Q/As of all states from common location
  - To compare data across all states

Discussion – a Paper Based on All Data?

- Contributions
  - Analysis of current situation, perspective on gaps and opportunities with NLP
  - Dataset
- Logistics
  - Target venue
  - People
  - Timeline

About Next Lecture – Lecture 19

Lecture 19 Outline

- Event extraction
- Spatio-temporal analysis

13	Oct 1 (Tu)	Language model – comparing arch, finetuning - **Guest Lecture**
14	Oct 3 (Th)	Language model – comparison of results, discussion, ongoing trends– **Guest Lecture**
15	Oct 8 (Tu)	PROJ REVIEW
16	Oct 10 (Th)	Using lang models to solve NLP tasks
17	Oct 15 (Tu)	QUIZ 2
	Oct 17 (Th)	
18	Oct 22 (Tu)	Entity extraction, linking
19	Oct 24 (Th)	Events extraction, spatio-temporal analysis
20	Oct 29 (Tu)	Topic Analysis