How To Supercharge Named Entities to Find and Use Relevant Digital Assets

Joseph Busch (aka Dr. Taxonomy)
Outline

- What is named entity recognition (NER)
- NER case studies
- Tools & resources
What is named entity recognition (NER)

- identifying named entities mentioned in unstructured text, and classifying them into pre-defined categories such as people, organizations, locations, events, and other things such as dates, time expressions, quantities, monetary values, percentages, codes, etc.

What is natural language processing (NLP)

After Mark Butler, CTO, Voise Inc.
What can NLP do

Categorize content: What is the content about?

Identify people, organizations, places, events, things: Identify and disambiguate entities (Kotonu is the French name of the capital of Benin; Apple the technology company vs. Apple the record company)

Identify grammatical relationships in sentence: Find subjects and objects of all verbs in the text

Extract actions: Identify related series of verbs (enter restaurant, sit down, read menu, order food, eat food, pay bill, leave)

Normalize disparate names for entities: Change “Kotonu” to long/lat - 6.3703° N/2.3912° E, “2.4 million” to 2,400,000, and “2006” to 1/1/2006-12/31/2006

Sets up capability to understand context, not just keywords.

After Mark Butler, CTO, Voise Inc.
Evaluation

- Recall – the number of true positive results divided by the number of all samples that should have been identified as positive.
- Precision – the number of true positive results divided by the number of all positive results, including those not identified correctly.
- F1 Score – a measure of a test's accuracy calculated from the precision and recall of the test.

https://commons.wikimedia.org/wiki/File:Precisionrecall.svg
NER use cases

- Which organizations were mentioned in the news article?
- Were specified products mentioned in complaints or reviews?
- Does the tweet contain the name of a person?
Case studies

Chronicle of Higher Education

Oracle Press Room
Entity recognition demo: Chronicle of Higher Education

- How can named entities be classified?
  - Government agency
  - Philanthropy
  - College or University

- Other named entity attributes for colleges and universities
  - Location
  - Size (students, employees)
  - Other criteria (e.g., Carnegie classification)
Named entity enrichment: Central Michigan University “Is…”

| instance of | public research university | 1 reference |
| public educational institution of the United States | 1 reference |

| located in the administrative territorial entity | Mount Pleasant | 1 reference |

| students count | 27,836 | 1 reference |

| employees | 2,500 | point in time September 2020 | 1 reference |

| Carnegie Classification of Institutions of Higher Education | doctoral university: higher research activity | 1 reference |
| high undergraduate | 1 reference |
| research doctoral: professional-dominant | 1 reference |
| four-year, large, primarily residential | 1 reference |
| four-year, full-time, selective, higher transfer-in | 1 reference |
| professions plus arts & sciences, some graduate coexistence | 1 reference |
Entity recognition demo: Oracle Newsroom

Press Release
University of Tennessee System Upgrades Finance and HR Tech with Oracle Fusion Cloud Applications

Statewide higher education system replaces SAP with Oracle Cloud to bring together finance and human resources to identify process improvements and improve decision-making

Issued: March 2, 2020

The University of Tennessee system has selected Oracle Fusion Cloud Applications to replace its on-premise SAP finance and HR systems, and also to replace the Oracle Fusion HRMS with Human Capital Management (HCM). The contract will help improve the learning and research environment for students and faculty, as well as enhance the quality of education provided.

The University of Tennessee system, which includes five public universities and two statewide institutions, makes more than $2.3 billion in capital investments per year. As part of its "Tennessee: Administrative Systems for Higher Education (TASH)" program, the UT System selected Oracle Fusion Applications for its ability to provide a complete view of operations and departments to its stakeholders. The TASH program also addressed the need for a strategic partner for the migration to Oracle Fusion Cloud Applications. Amongst the fields with which the UT System project team selected members to help facilitate transition to the new system, one of the most important for the finance department is process optimization.

"It is critical for universities to find ways to serve faculty and students more efficiently. Cloud-based systems can help campuses do that by improving and automating business processes," said UT President Randy Boyd. "We are confident that the Oracle infrastructure, bolstered by Oracle Fusion Applications, will enhance operational efficiency and position the UT System for continued innovation.

With Oracle Cloud IaaS and Oracle Cloud HCM, University of Tennessee System leaders will be able to gain real-time visibility into their business processes, improve efficiency, and provide a better experience for employees. With quarterly upgrade cycles, Oracle Fusion Applications will also give the UT System the freedom to access the new features that are added every 90 days, without disruption to business operations.

"With Oracle Fusion Applications, the University of Tennessee System will be able to centralize and simplify finance and HR processes to keep its operations in line," said Murphy Sisler, group vice president of product development, Oracle. "Every bit of efficiency gained from these solutions is crucial to dedicated faculty and university's commitment to serving students."
Oracle top-level taxonomy architecture

Oracle Cloud Applications
- Enterprise Performance Management
- Government & Education

“Is a” groups of Products

Oracle Cloud Infrastructure
- Cloud Regions
- Data Centers
- Product Line
- Technology
- Application
- Industry Solution

Oracle top-level taxonomy architecture

Oracle Cloud Infrastructure
- Compliance
- Container and Functionality
- Oracle E-Business Suite
- Oracle Database
- Oracle Fusion Applications
- Oracle Fusion ERP

Oracle Cloud Applications
- Cloud Infrastructure
- Fusion Applications
- Fusion ERP

Oracle Cloud Infrastructure
- Cloud Regions
- Data Centers
- Product Line
- Technology
- Application
- Industry Solution

Oracle Cloud Applications
- Enterprise Performance Management
- Government & Education

“Is a” groups of Products

Oracle Cloud Infrastructure
- Cloud Regions
- Data Centers
- Product Line
- Technology
- Application
- Industry Solution

Oracle Cloud Applications
- Enterprise Performance Management
- Government & Education

“Is a” groups of Products

Oracle Cloud Infrastructure
- Cloud Regions
- Data Centers
- Product Line
- Technology
- Application
- Industry Solution

Oracle Cloud Applications
- Enterprise Performance Management
- Government & Education

“Is a” groups of Products

Oracle Cloud Infrastructure
- Cloud Regions
- Data Centers
- Product Line
- Technology
- Application
- Industry Solution

Oracle Cloud Applications
- Enterprise Performance Management
- Government & Education

“Is a” groups of Products
What about non-text media?

**Voice to text**

**Surrounding text**

https://www.washingtonpost.com/food/2022/04/09/matzoh-balls-floaters-recipe/
Tools

- **Toolkits**
  - Apache Open NLP machine learning based toolkit for the processing of natural language text
  - GATE open source software toolkit
  - NLTK Natural Language Toolkit for building Python programs to work with human language data
  - SpaCy natural language processing in Python
  - Stanford NER Java implementation of a Named Entity Recognizer

- **Applications**
  - Cogito provides named entity recognition annotation services to extract entities from text creating NLP training data for machines
  - Expert.AI applies artificial intelligence to natural language understanding of text
  - NetOwl entity extraction in multiple languages using NLP and machine learning
  - PoolParty semantic technology platform to enrich content by combining artificial and human intelligence
Resources

Questions

Joseph Busch
jbusch@taxonomystrategies.com
mobile +1-415-377-7912
Overview

There’s lots of software available that does a good job of identifying named entities – the names of people, organizations, events, places, etc. – that occur in text. Too often these are just used as keywords to find digital assets without any further differentiation. For example, an organization could be a public company, U.S. government agency, an institution of higher education, or something else. Being able to identify that a named entity is a specific type of organization can be important in determining whether a digital asset is relevant to a particular search. It can also be useful in adding further context to the digital asset. For example, if we know that a named entity is an institution of higher education, we could further differentiate it by size and location, and even link to a short profile of the organization in Wikipedia or to the institution’s website itself. This talk explains how you can easily build up metadata related to named entities to improve search so you can accurately find and use relevant digital assets, and provides real world examples from client projects.