Semantic Mobile Search
Andreas Kaltenbrunner
Senior Director of Data Analytics
NTENT

- Semantic search technology
- Born from the merge of Convera and Firstlight ERA in 2010
- Convera is previously known for enterprise search engine RetrievalWare and the web-scale semantic search engine Excalibur
- Offices in:
  - New York, NY (HQ)
  - Carlsbad, CA (Engineering)
  - Barcelona, CAT (R&D)
  - Vienna, VA (USA)
  - London (UK)
At-A-Glance: NTENT’s White-Label Platform

NTENT-Powered Applications

NTENT Technology Core
- Natural Language Processing
- Query Understanding
- Knowledge Base
- Intent Signals
- Applied Data Science
- Question and Answer Engine

NTENT Skill & Capabilities
- Location Based Search
- Questions & Answers
- Points of Interests
- Weather
- Sports
- Wikipedia
- Shopping
- Food
- Travel
- News
- Command & Control
- Movies

NTENT Platform

Data Analytics

Voice Assistant

Web Search

Mobile Browser

Ad Monetization
AGENDA

- Semantic Resources
- Document Processing
- Experts
- Query Understanding
- Usage Data Analysis
SEMANTIC RESOURCES
KNOWLEDGE RESOURCES

Composed of Four Parts

• Ontology (LI)
• Lexicon (LD)
• Onomasticon (mostly LI)
• Linguistic Rules (SemReps, Case Frames)
LEXICON

• Language **specific**
  (one lexicon per language)
• One or more expression associated with each concept from the ontology
• Includes expression- specific attributes
• Current languages:
  • English
  • Russian
  • Turkish

ONOMASTICON

- Generally Language **independent**

- One or more expression associated with each entity

- Includes expression-specific attributes

- Sometimes not language independent e.g. city of Geneva in Switzerland:
  - Geneva (en)
  - Genève (fr)
  - Ginebra (es)
  - Ginevra (it)
  - Genf (de)

- **Generic People**

- **Intermediate Singer**

- **Specific Lady Gaga**

- **en: Lady Gaga**
  - **fr: Lady Gaga**
  - **es: Lady Gaga**
LINGUISTIC RULES
VARIOUS KINDS: CASE FRAMES, SEMREPS, ETC.

-- Given a context of lunch or dinner
-- and a cuisine is selected
-- returns list of restaurants
If
{
  Sequence
  {
    Or
    {
      Concept['s:men.00BBB', desc = "dinner", match = "self"],
      Concept['s:men.01A90', desc = "lunch", match = "self"]
    }
  }
}

Templates
{
  -- Covers questions such as "list of N ...
  If
  {
    Sequence {
      Optional[ Text["give me"] ],
      Optional[ Or[ Text["a"], Text["the"] ] ],
      Text["list"],
      Optional[ Text["of"] ],
      ZeroOrMore { Numbers_Entity, variables = "answer_count" },
    }
    Interpretation{
      name = "List of things", id="LISTOF",
      -- specify how many answers to fetch.
      FieldQuery[ name='Common максAnswerCount', Variable["answer_count"]]
    }
  }
}
DOCUMENT PROCESSING
SEMANTIC ENGINE

Semantic Engine

Ontology / Lexicon / Onomasticon

Concepts (common nouns)
Entities (proper names)
SEMANTIC ENGINE FUNCTIONS

- Language detection
- Boilerplate detection and removal
- Tokenization and lemmatization
- Part of speech tagging
- Morphological analysis
- Entity extraction and general dictionary matching
- Concepts identification and disambiguation
- Document scoring and classification
Rockets can't afford to have Dwight Howard slowed by injury
ENTITY EXTRACTION

Attested Entities

• Named Entities: people, places, organizations, brands, works of art, etc.
  • Represented in the ontology with relations to other concepts

Unattested (not in the Onomasticon)

• Named Entities: people, places, organizations, etc.
  • Inferred people, places, organizations, etc. based on clues (e.g., “Mayor”, “Hospital”, “Inc. “, “Mr.”)
  • Machine learning based NER
• Other Entity Types: phones, monies, date/time, etc.
  • Pattern based, localized
IDENTIFY DOCUMENT CHARACTERISTICS

Uses rule-based and machine-learning classification

**Example:** If document mentions a sports team or an athlete, then classify as sports-category

- Identify broad document topic (e.g., health, politics)
- Identify aspect around a topic (e.g., entertainment, places to stay, restaurant, weather)
- Identify document genre (biography, FAQ, contract, consumer guides, recipes, analyst opinion)
- Identify spam, adult, offensive content, smoking, weapon use, gambling, unfortunate events
- Identify redaction quality (scientific article, conversation)
- Identify special document characteristics (many links, typos, shopping cart, login)
EXAMPLE: IAB category classifier

Classification of queries into predefined categories

IAB (Interactive Advertising Bureau) topic taxonomy
https://www.iab.com/guidelines/taxonomy/

- Takes advantage of rich, pre-trained word embedding models that leverage large corpora.
- No need to annotate large amount of queries
- Instead, generate suitable keywords for each of the categories.
- Predict the category in the following way:
  1. Generating a 'canonical vector', encodes the overall semantic properties across the keywords of a category.
  2. Assign an uncategorized query to closest 'canonical vector',

<table>
<thead>
<tr>
<th>Category cohort</th>
<th>2018 September</th>
<th>2019 January</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology and Computing</td>
<td>13.4%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Pop Culture</td>
<td>9.7%</td>
<td>11.9%</td>
</tr>
<tr>
<td>Travel</td>
<td>4.8%</td>
<td>6.3%</td>
</tr>
<tr>
<td>News and Politics</td>
<td>7.5%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Television</td>
<td>4.7%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Sports</td>
<td>2.6%</td>
<td>3.9%</td>
</tr>
<tr>
<td>Shopping</td>
<td>3.2%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Events and Attractions</td>
<td>2.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Home and Garden</td>
<td>2.8%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Food and Drink</td>
<td>3.2%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Movies</td>
<td>4.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Family and Relationships</td>
<td>2.3%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Education</td>
<td>2.7%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Automotive</td>
<td>1.9%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Religion and Spirituality</td>
<td>2.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Pets</td>
<td>1.5%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Healthy Living</td>
<td>1.9%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Video Gaming</td>
<td>1.3%</td>
<td>2.0%</td>
</tr>
<tr>
<td>ADULT</td>
<td>13.7%</td>
<td>13.9%</td>
</tr>
<tr>
<td>Fine Art</td>
<td>1.4%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Business and Finance / Business</td>
<td>1.8%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Business and Finance / Economy</td>
<td>2.5%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Science</td>
<td>1.3%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Music and Audio</td>
<td>1.5%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Business and Finance / Industries</td>
<td>1.5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Style and Fashion</td>
<td>1.5%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Careers</td>
<td>1.3%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Medical Health</td>
<td>1.0%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Real Estate</td>
<td>0.9%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Hobbies and Interests</td>
<td>0.8%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Books and Literature</td>
<td>0.6%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Personal Finance</td>
<td>0.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Impact of an adult filter on query category distribution
MOBILE SEARCH RESULTS - BEYOND TEN BLUE LINKS
APPLICATION: EXPERT ANSWERS

Query

PF Chang near Encinitas

Formal Interpretation

1. Interpretation for 'MEAL_PLAN' (rank:1.14167)
   1. Situation (id:MEAL_PLAN)
      1. Action (id:RESTAURANT_ACTION_EST, name:Let's go out for a meal, model:null, rank:1.25)
         MEAL_PLAN access P.F. Chang's Encinitas, CA
         - FieldValue: Query.Class = local (null)
         - FieldValue: Query.Topic = Foodstuff Vertical (vt:vt.CDSW9)
         - FieldValue: Directory.Business.generalCategory = restaurant (s:pst.001CR)
         - FieldValue: Directory.Business.name = P.F. Chang's (s:pst.00KZ3)
         - FieldValue: Common.Location.city = Encinitas, CA (s:usg.1652705)
         - FieldValue: Common.Location.region = California (s:usg.0008D)
         - FieldValue: Common.Location.regionCode = CA (null)
         - FieldValue: Common.Location.country = United States of America (s:geon.00CQ)
         - FieldValue: Common.Location.countryCode = US (null)
         - FieldValue: Common.Location.lat = 33.037 (null)
         - FieldValue: Common.Location.long = -117.292 (null)
EXPERTS – BEST SOURCES OF KNOWLEDGE

A single authority on a particular type of result or subject type. Examples:
  Yelp = Local Business Listings
  Stats.com = Major league sports data

NTENT treats its own ingested results as experts:
  NTENT news and web-based search index

Three flavors of expert:
  External API call – an external API is called that resides outside of NTENT data centers
  Ingested data from third parties – offline or batch data is regularly ingested
  NTENT’s own expert data
QUERY UNDERSTANDING
USAGE DATA
ANALYSIS
NTENT Data Analysis, Transformation and Activation

Our deterministic data is very valuable, can be monetized and used to enhance digital marketing and product development.

- Collect
  - NTENT-powered applications & devices
  - Search Term
  - Query Category
  - Related Queries
  - Trending Queries
  - Timestamp
  - IP Address
  - Lat/Long

- Analyze
  - Device ID
  - Ad Clicks
  - Device Details
  - User Agent
  - Organic Clicks
  - Widget Interactions
  - Tab Interactions

- Transform
  - Search & Content Expertise
  - Semantic Engine
  - Interpretation Engine
  - NTENT Answer Engine
  - Machine Learning & Analytics
  - Natural Language Processing & Ontology

- Integrate
  - API Integrations
  - Behavioral Insights
  - Audience Segmentations
  - Ad Target & Retargeting
  - Attribution
  - Audience Mapping
  - Product development
  - Pricing

NTENT-powered applications & devices

NTENT - powered applications & devices
Example: Churn and Activity Prediction

Objectives:

- Create classifiers to predict user activity and possible user churn.
- Detect positive or negative impact of certain features on user retention.
Example: Real Time Query Analysis

Objectives:

• Real time monitoring and visualization systems for search traffic analysis.
• Detect trends and incidents
Objectives:

- Improve search results through location and trajectory analysis.
- Target advertising based on movement data.