Entity Disambiguation: Our Approach

At Finch Computing, we build new ways of interacting with information. Perhaps nowhere is this more apparent than in our text analytics solution, Finch for Text® which makes human-generated text machine-readable. We say Finch for Text® is "software that reads and reasons" because proprietary technologies in the product enable it to extract, disambiguate and enrich entities and to assign sentiment to these entities in ways other solutions just can't replicate. Below is a sampling of how we approach entity disambiguation in particular in support of a number of business and mission-critical use cases.

What is Entity Disambiguation?

Entity disambiguation refers to the ability to resolve an entity's identity to a knowledgebase. It is not merely entity-type classifying – as in determining that a reference to "George Washington" is a reference to the person and not the bridge in New York City, for example. Instead, entity disambiguation involves correctly distinguishing between two identically named entities of the same type – as in, John Roberts the Chief Justice of the U.S. Supreme Court, John Roberts the Fox News correspondent, or any one of the hundreds of individuals in the world named John Roberts.

Doing this requires understanding entities and their surrounding context. In involves developing technology that is capable of reading text as a human would – understanding nuance and inferences, nicknames and short-hands, misspellings and variances in upper and lower case text. It also involves the development and curation of massive knowledgebases, full of entities and knowledge *about those entities* in order to correctly interpret their surrounding context in a piece of text.

Many text analytics solutions will claim to do entity disambiguation, but many either use a modified definition of the practice, or concentrate only on a particular domain or entity type; even worse, they deliver accuracy that is sub-par at best.

Our Knowledgebases

At Finch Computing, we have developed huge knowledgebases of people, places and organizations. These include hundreds of millions of discrete entities and make our knowledgebases incredible assets. However, what's of the most value is the knowledge we have, as referenced above, *about* those entities – knowledge of the topics associated with an entity, facts about an entity, key phrases that often appear with or near and entity.

Using the John Roberts example above, our knowledgebase entry for the Chief Justice would include terms like law, judge, legal, courts, rulings, etc. These terms comprise a numeric, machine-readable "topic" that is linked to the John Roberts entity in our knowledgebase. Key phrases like "the nine justices" or "in a historic ruling" are similarly linked to the entity. Facts about Justice Roberts – like where he went to school, when he was born, past positions he's held – are also stored in our knowledgebases.

All of this information – billions of pieces of knowledge – is immediately at our disposal when analyzing a piece of unstructured text. We constantly add to our knowledgebases with human and machine-driven updates. And we can add customer-specific information to it with ease.

Entity Disambiguation: Our Approach

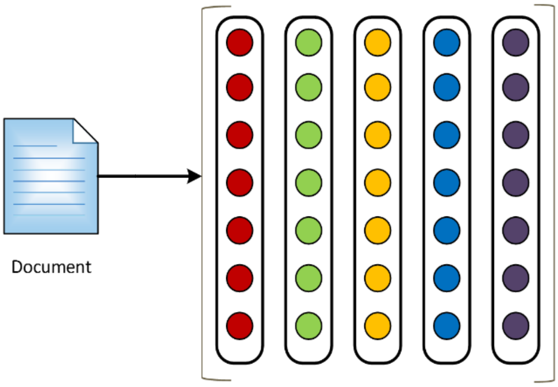
A Unique, Proprietary Approach

Finch for Text® takes a proprietary, context-based approach to entity disambiguation. We hold patents that involve improving traditional, algorithmic approaches in order to assess multiple, more dimensionalized components of a document. As such, Finch for Text® can understand entities, key phrases, topics and embedding vectors in order to link an entity appearing in text to our knowledgebase. This essentially creates a mathematical feature model for every piece of text and allows us to quickly and intimately understand the entities referenced in it.

Powered By an In-Memory Analytics Engine

Finch for Text®'s ability to perform accurately, and at high speeds and high volumes, is enabled by an underlying in-memory, predictive analytics platform called FinchDB®. At its core, FinchDB® is a JSON, doc database. It can handle thousands of disambiguation queries every second. One of the most distinctive qualities of FinchDB® is its ability to embed models in queries – including predictive models. This means users get real-time, accurate query results, even as their data is changing.

We turn documents into feature models in order to understand entities and their associations in entirely new ways.



The document multi-component feature vector (document feature model) includes:

- **Entities:** Via Entity Extraction Models
- **Disambiguated Entities:** Via Entity Disambiguation Models
- **Key Phrases:** Extracted via Key-Phrase Extraction Models
- **Topic Vectors**: Inferences detected via Topic Model
- **Embedding Vectors:** Neural Embedding Vectors via Embed Representation Models

How can it be used?

Entity disambiguation supports a host of research, intelligence and analysis use cases across the commercial and federal spaces. We currently support customers in these domains and more, and would be happy to speak with you about your need to understand and disambiguate entities in unstructured text.

To learn more, please contact us at sales@finchcomputing.com or visit us at www.finchcomputing.com.