Word Embeddings for Fuzzy Matching of Organization Names

Interesting piece from Basis Technology

fraXses utilises Basis Technology's Rosette for the management of unstructured data. Rosette's name matching is enhanced by word embeddings to match based on semantics as well as phonetics

Tracking mentions of particular organizations across news articles, social media, and internal communications are integral to the workflow of dozens of use-cases across industries. However, it can be especially challenging to match names of companies and organizations because of misspellings, abbreviations and names in non-Latin scripts. For organization names, words like "corporation," "group," or "incorporated" are often used interchangeably, leading to missed matches when deduplicating databases, or searching for specific corporate entities. Also, people are likely to forget if it was the "London Philharmonic Orchestra" or the "London Philharmonic Symphony."

To solve this problem we've added text embeddings, one of the most powerful results of current deep learning research, to our name matching algorithm. This allows Rosette to match entities based on words with similar meanings, rather than only phonetics. For example, a search for "Eagle Drugs, Inc." of the NASDAQ database will fuzzy match "Eagle Pharmaceuticals, Inc." because "drugs" and "pharmaceuticals" are close in meaning.

Slow down, what's fuzzy name matching again?

Traditional name matching methods look at names as sequences of letters, and generate an exhaustive list of name variations to check against each name presented for matching. The process requires continual generation and storage of name variations and only accommodates names written in Latin-based alphabets.

Rosette uses machine learning-based name matching, built over patterns of language and cultural use. Fuzzy name matching accommodates names in non-Latin scripts by applying statistical knowledge of how names vary and how each letter or group of letters sounds in different languages, increasing accuracy and speed.

Rosette's name matching recognizes and addresses 13 different name-related phenomena such as phonetic similarity, nicknames, and transliteration spelling errors.

How do word embeddings improve match scores?

Text embeddings represent text with numbers—as vectors. Calculating the embeddings of two documents, phrases or words lets you evaluate how semantically similar they are based on context, content, and associations. (Learn more about how text vectors work on our previous blog posts: part I and part II.)

We've incorporated text embeddings into the "second pass" of our name matching software, the part that refines the similarity scores of names that Rosette has already identified as probable matches. The addition of text vectors increases accuracy without falsely identifying every bank, trust, or credit union as the same business.

For example, take State Street Bank and State Street Corporation. Looking solely at phonetics, Rosette first recognizes that "state" and "street" both match exactly, but "bank" and "corporation" do not, leading to a similarity score of approximately .70.

However, while they aren't synonyms, "bank" and "corporation" are still semantically similar (at least more so than say, "bank" and "llama"), which means they have similar vector representations. Text embeddings in our name matching process utilizes that similarity, raising the similarity score to approximately .76.

Don't let your matches get lost in translation

Often the direct translation of an organization's name is not a perfect match. This is especially true for divergent languages like Spanish and Korean as opposed to Spanish and Italian which share romance roots. Enter text embeddings to improve match scores for cross-lingual names that might otherwise be missed.

For example, "United Nations" in Japanese is represented 国際連合, however, the characters 国際 actually translate to "international," which isn't the same as "nations." Because both "nations" and "international" are in a similar vector space, text embeddings ensure that 国際 correctly matches with "nations." Similarly, Honda Motor Company is written 本田技研工業株式会社. 工業 is Japanese for industry, which isn't normally a synonym for motor, but with text embeddings, the two are correctly matched.

Phonetics and semantics: a text analytics power couple

Rosette's name matching evaluates both phonetic and semantic similarity to determine matches, a powerful combination designed to optimize both accuracy and precision.

A system that only identifies phonetic similarities may still connect most related entities, but with lower recall, as in the State Street example above. Alternatively, a system that relies solely on semantic similarity and disregards phonetics may mistakenly match Citizens Financial Group with Bank of America, two distinct corporate entities. Our approach brings these two techniques together, providing our customers with the even higher recall and precision.