Task: Write a description of how you would find references to the company currently known as **Broadcom Inc.** at any point in time since 2008. Describe your choices and how you found the company's names over time.
Hint: *companies may change names for several reasons, such as restructurings, mergers, demergers, acquisitions, etc.*

To track references to the company currently known as Broadcom Inc. over time since 2008, I would employ a combination of techniques, including data mining, natural language processing (NLP), and historical research.

Note: I provide 2 solutions: an ML solution leveraging key NLP concepts, and a deep learning solution leveraging LLMs.

Solution 1 - ML Solution

Assumptions

1. All articles containing information about company name changes should be available in order to make an exhaustive timeline.
2. Exact dates or temporal phrases should be present in atleast some articles to avoid conflicts, e.g.
    a. In 2015, Company A got acquired by B. In 2015, B merged with C to form D. - Ambiguous case.
    b. In **early** 2015, Company A got acquired by B. **Later that same year**, B merged with C to form D. - Valid case.

Implementation

1. **Data Collection:**

1. **Gather a comprehensive dataset** consisting of news articles, press releases, financial reports, SEC filings (such as 10-K and 10-Q forms), investor presentations, and other relevant documents spanning the period from 2008 to the present.
2. This dataset should cover a wide range of sources to ensure a comprehensive understanding of how the company has been referred to over time.

2. **Identification of Key Corporate Events**:

1. Utilize historical records, company filings, and news archives to **identify significant corporate events** such as mergers, acquisitions, rebranding, and restructuring activities involving Broadcom Inc. and its predecessor entities during this timeframe.
2. Specifically, focus on dates and temporal phrases when the company changed its name or underwent significant corporate transformations. For example:
    a. In **2015**, Avago Technologies **acquired** Broadcom Corporation and adopted the Broadcom Inc. name for the combined entity.
    b. **Prior to the acquisition**, Broadcom Corporation was known simply as Broadcom.
3. Methods such as semantic-matching of keywords (mergers, acquisitions,, etc.) with each article using word embeddings can be used.

3. **Alias Generation:**

1. Using the retrieved set of key corporate events articles, develop algorithms or rulesets to **generate a list of aliases or alternative names** associated with the company at different points in time. This includes variations due to mergers, acquisitions, subsidiaries, rebranding, abbreviations, and any other corporate actions.
2. For example, for Broadcom Inc., aliases may include:
    a. Broadcom Corporation
    b. Avago Technologies
    c. Avago/Broadcom
3. Employ NLP techniques such as named entity recognition (NER) to identify references to the company within the text corpus.

4. **Entity Resolution**:

1. Apply entity resolution methods to match identified aliases with mentions of the company in the documents. This involves **linking each mention of the company to the appropriate timeframe** based on the known dates of name changes and corporate events.
2. This involves:
    a. Standardization of the dataset by mapping known aliases to the company name.
    b. Duplicate detection and tagging using fuzzy matching. Match articles using aliases and key corporate events. As a result, we have a set of (alias, corporate event, timeframe) where the duplicates will be identified.
    c. Conflict resolution and merging. Deduplicate sets to have 1 alias per company event and timeframe.

5. **Temporal Analysis**:

We can also **integrate temporal tagging** for ensuring consistency to facilitate chronological analysis and visualization. Assign each mention of the company to the corresponding timeframe to track its evolution over time accurately.

7. **Validation and Iteration**:

Validate the accuracy and completeness of the extracted references through manual verification and comparison with reliable sources. Iterate on the process to refine algorithms and improve the quality of results.

Evaluation Metrics

1. **Precision**: Evaluate the number of correct aliases retrieved out of all possible aliases.
2. **Accuracy:** Evaluate the number of correct (alias, timeframe) pairs out of the entire company history.
3. **F1-score:** A high f1-score is desirable for evaluating (alias, timeframe) pairs.

Advantages

1. Handles unstructured data.
2. Handles ambiguous references to the company.
3. Handles conflict resolution and noisy data.

Limitation

1. If not articles from a particular timeframe are present, there will be missing values.

Solution 2 - Retrieval Augmentation

We can also leverage the use of powerful LLMs and techniques such as RAG to automate the above process.

Assumptions

3. All articles containing information about company name changes should be available in order to make an exhaustive timeline.
4. Exact dates or temporal phrases should be present in atleast some articles to avoid conflicts, e.g.

a. In 2015, Company A got acquired by B. In 2015, B merged with C to form D. - Ambiguous case.
b. In **early** 2015, Company A got acquired by B. **Later that same year**, B merged with C to form D. - Valid case.

Implementation

1. **Data Collection:** same as above.

2. **Knowledge Base Integration with LLM:**

1. Use the entire dataset collected as a knowledge base for the company's history since 2008.
2. Create a vector database and integrate it with an instruct-tuned LLM.
3. This can be done using PineCone (for creating index DBs) and langchain (for LLM integration).

3. **Prompt Engineering:**

1. Break down the complex task of finding company references with their corresponding year into smaller, simpler subtasks.
    a. This aids the LLM to rationally arrive at the final solution.
    b. It also aids in debugging intermediate steps of LLM's reasoning capabilities.
2. Prompt the LLM using **chain-of-thought** to do the following subtasks:
    a. Identification of key corporate events
    b. Alias generation
    c. Entity resolution
3. The LLM retrieves relevant documents in the vector database and provides answers without needing to further fine-tune on company data.

4. Validation and Iteration: same as above.

Evaluation Metrics

4. **Precision**: Evaluate the number of correct aliases retrieved out of all possible aliases.
5. **Accuracy:** Evaluate the number of correct (alias, timeframe) pairs out of the entire company history.
6. **F1-score:** A high f1-score is desirable for evaluating (alias, timeframe) pairs.

Advantages

4. Handles unstructured data.
5. Handles ambiguous references to the company.
6. Handles conflict resolution and noisy data.

Limitation

2. If not articles from a particular timeframe are present, there will be missing values.

By following Any of the provided approaches, we can effectively track references to Broadcom Inc. (and its predecessor entities) in textual data over time, providing valuable insights into the company's historical evolution.