Semantic Frame Identification Using the Natural Language Toolkit and Wordnet

Vishnu Datta Jayanti
Senior at Allen High School
Allen, Texas
vishnudattaj@gmail.com

Abstract—Semantic Frame identification plays a critical role in understanding the deeper meaning of natural language beyond the surface level. This paper presents an approach using the Natural Language Toolkit (NLTK), WordNet, and FrameNet to identify potential semantic frames incorporated within a corpus. The project utilizes text normalization, word and sentence tokenization, part-of-speech tagging, and named entity chunking to identify lexical units and frame elements. A custom-made gazetteer was designed to correct named entity recognition (NER) errors, consequently improving frame element detection. This approach demonstrates potential for scalability, especially in the financial world.

I. Introduction

The corpus selected for this study is *Reminiscences of a Stock Operator* by Edwin Lefèvre (1923), a world-renowned work in financial literature that traces the journey of a stockbroker navigating the stock market. Given the thematic emphasis on the stock market, commercial exchange, and organizations, three semantic frames were selected from FrameNet: *Capital_stock*, *Commercial_transaction*, and *Businesses*. These frames were chosen because of the narrative's rich content regarding the corporate world and Wall Street, making the extraction of lexical units and frame elements more straightforward.

II. Preprocessing

A central component of this project involves accurately identifying named entities—particularly people, organizations, and monetary values—to support frame element (FE) extraction. (NLTK) provides a built-in named entity chunker that accurately identifies named entities in general contexts, but struggles with domain-specific texts. For instance, the abbreviation "UP", which stands for the fictional organization *Union Pacific Company*, was mistakenly labeled as a generic proper noun rather than an organization. To address these inconsistencies, selected portions of the corpus were manually reviewed to locate potential (NER) misclassifications. These corrections were manually stored in the plain-text file `gazetteer.txt`, which associates misidentified entities with the proper labels. During the preprocessing phase, the program checks the entities extracted from each sentence and refers to the gazetteer to update the entity list accordingly. This refined list is essential to ensure the success of (FE) extraction.

III. Workflow

Each semantic frame used in this project incorporates two types of functions: an identifier and a matcher. The identifier determines whether a given sentence contains an instance of a semantic frame by iterating through each word in the sentence and comparing it with the frame's lexical units using both *Wu-Palmer* and *Leacock-Chodorow* similarity. These measures help determine whether the lexical units are present within the sentence. If the word scores highly in both metrics, the function returns `True` along with the word representing the lexical unit. This indicates that the semantic frame is present within the sentence. While identifier functions follow a consistent structure, matcher functions widely differ based on the semantic frame. Because each frame contains a unique set of (FE)s, each with its own set of criteria, each matcher function contains distinct logic.

IV. Capital Stock

FrameNet identifies five frame elements for *Capital Stock*: the issuer, shareholder, stock, type of stock, and amount. The issuer can be determined by identifying entities labeled as *ORGANIZATION* within the sentence. Finding the shareholder requires more nuanced detection as the shareholder ranges from pronouns to *PERSON* entities. Possessive constructions (e.g., apostrophes such as *John's*) are prioritized, followed by possessive pronouns (e.g., *his*, *her*), and finally by entities tagged as *PERSON*. There can only be one shareholder. The stock element detection is

similar to the logic used for semantic frame identification, as each word is compared with WordNet's synsets for stock and shares. If a positive match is calculated, the word is determined to be the stock element. The type of stock is typically absent; however, in rare instances where the phrases "preferred stock" or "common stock" occur, the type is assigned accordingly. The amount is identified by locating all cardinal numbers within the sentence and tracking them within a list.

V. Commercial Transaction

Similarly, for Commercial Transaction, FrameNet identifies five frame elements for Commercial Transaction: the buyer, seller, goods, price, and currency. The program begins by determining whether the sentence represents a "buy" or a "sell" by iterating through each word and matching it with WordNet synsets using *Wu-Palmer* and *Leacock-Chodorow* similarities. Once detected, the buyer and seller are determined based on their approximate position from the "buy"/"sell" word. Once the buyer or seller is determined, the program scans the sentence for the presence of the word "from" or "to". If an entity (*PERSON* or *ORGANIZATION*) appears immediately after the word "from" in a sentence involving a purchase (and the sentence is labeled as a "buy"), it is classified as the seller. Conversely, if an entity appears near the word "to", it is classified as the buyer. This approach enables the identification of buyer and seller entities even in the absence of explicit triggers such as "buy" or "sell". The process of determining the goods incorporates both *Wu-Palmer* and *Leacock-Chodorow* similarity as well, this time by comparing words to WordNet's synset for commodity. To extract the amount, a custom entity label, *MONEY*, was introduced. The *MONEY* entity was determined by finding all instances of consecutive cardinal digits (including the conjunction "and") that contain either a $ sign or the word "dollar". The gazetteer is significantly important in this situation, as (NER) mistakenly identifies the numbers ten through ninety-nine as adjectives, as opposed to cardinal digits. Because the corpus exclusively references American currency, the unit is set to dollars by default if a *MONEY* entity is determined.

VI. Businesses

For the Business frame, FrameNet identifies seven different (FE)s; however, only three of them are utilized for this project. The three utilized are the business, the descriptors, and the location. The business is identified by extracting all entities labeled as *ORGANIZATION* within the sentence. Once a business entity is detected, its descriptor is determined by locating adjectives within two words of the organization name. The location is extracted by identifying entities labeled as *GPE* (geopolitical entity), which typically denote cities, states, or countries.

VII. Conclusion

This project demonstrates a practical approach to semantic frame identification using (NLTK), WordNet, and FrameNet. By utilizing text normalization, word and sentence tokenization, part-of-speech tagging, named entity chunking, and lexical units, the project effectively detects semantic frames and extracts frame elements from the sentence. This methodology shows promise in extracting similar information from other financial documents.