Coreference Feature Extraction: Technical Specification

1. Chain Diversity Entropy

How It Is Computed:

For each coreference chain (i.e., a group of mentions that refer to the same entity), we calculate the distribution of mention types: entity, pronoun, and definite. We then compute the Shannon entropy:

$$H = -\sum_{i} p_i \log_2(p_i)$$

where p_i is the proportion of mentions of type i within the chain. The final feature is the average entropy across all multi-mention chains.

What It Represents:

Entropy measures the diversity of referential expressions. A chain using only one type of mention (e.g., only pronouns) has low entropy. A chain that uses a mix (e.g., names, pronouns, noun phrases) has high entropy.

Real-Life Correspondence:

In natural writing, we often vary how we refer to the same entity. For example, âĂIJDr. SmithâĂİ, âĂIJheâĂİ, and âĂIJthe researcherâĂİ may all refer to the same person. High entropy reflects this linguistic variety and coherence.

2. Chain Length Variance

How It Is Computed:

For each coreference chain (including singleton mentions as their own chains), we compute its length. Then we compute the variance across chain lengths:

$$\mathrm{Var} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

where x_i is the length of the i-th chain and $\bar{x}$ is the mean chain length.

What It Represents:

It shows how unevenly reference is distributed. Some entities might be mentioned many times, others just once. High variance reflects asymmetry in how frequently entities are discussed.

Real-Life Correspondence:

Important entities (like a main character or company) tend to appear more frequently than secondary ones. Texts with clear focus tend to have higher variance.

3. Long-Range Coreference Ratio

How It Is Computed:

For each multi-mention chain, we measure the sentence span between the first and last mention. If the span is ≥ 2, the chain is considered long-range. Then we compute:

$$\text{long\_range\_coref\_ratio} = \frac{\#\text{long-range chains}}{\#\text{multi-mention chains}}$$

What It Represents:

It measures how often entities are mentioned across multiple sentences. A higher value means the text maintains coherence over longer spans.

Real-Life Correspondence:

Long-range references appear in stories, essays, and reports, where the same person or object appears repeatedly. This is rare in short or shallow texts.

4. Mention Complexity

How It Is Computed:

We calculate the number of tokens (words) in each mention, then average across all mentions:

$$\text{mention\_complexity} = \frac{\text{total tokens in mentions}}{\text{total number of mentions}}$$

What It Represents:

Longer mentions suggest more descriptive, specific, or technical references. Short mentions are typically pronouns or generic terms.

Real-Life Correspondence:

Formal or technical writing may include complex phrases like âĂIJthe newly appointed director of engineering,âĂİ while informal writing may stick to âĂIJheâĂİ or âĂIJtheyâĂİ.

5. Singleton Ratio

How It Is Computed:

Mentions not connected to any other are considered singletons. The ratio is:

$$\text{singleton\_ratio} = \frac{\#\text{singleton chains}}{\text{total number of chains}}$$

What It Represents:

It tells us how many mentions do not appear again. A high value suggests the text introduces many unique or one-off entities.

Real-Life Correspondence:

Texts like lists, search engine results, or chats often have high singleton ratios. In contrast, coherent narratives tend to re-use key entities.

6. Pronoun Ratio

How It Is Computed:

We count all mentions of type pronoun, then divide by total mentions:

$$\text{pronoun\_ratio} = \frac{\#\text{pronoun mentions}}{\text{total mentions}}$$

What It Represents:

This measures how much the text relies on pronouns instead of full names or phrases. ItâĂŹs a signal of context dependence.

Real-Life Correspondence:

Spoken language and fiction tend to use more pronouns (e.g., âĂIJheâĂİ, âĂIJsheâĂİ, âĂIJitâĂİ), whereas formal reports use names or titles more often.

7. Chain Connectivity

How It Is Computed:

For each chain, compute:

$$\text{coverage} = \frac{\text{token span of chain}}{\text{total tokens in document}}, \quad \text{density} = \frac{\#\text{mentions in chain}}{\text{token span}}$$

$$\text{connectivity} = \text{coverage} \cdot \log(1 + \text{density})$$

The final value is the average connectivity across all chains.

What It Represents:

Connectivity captures how well-distributed and densely packed a chain is within the text. Higher values mean that mentions of the same entity are frequent and spread across the document.

Real-Life Correspondence:

This often occurs in detailed discussions, long articles, or essays where important entities appear multiple times throughout the text.

8. Average Cluster Size

How It Is Computed:

Only considering multi-mention chains, we compute:

$$\text{avg\_cluster\_size} = \frac{\sum_{i=1}^{K} \text{length of chain}_i}{K}$$

where K is the number of chains with more than one mention.

What It Represents:

This shows the average number of references per entity. Higher values suggest better tracking of key actors or objects.

Real-Life Correspondence:

In essays or fiction, important characters are referenced repeatedly. This leads to higher cluster sizes. In shallow texts, entities may only be mentioned once or twice.