

Annotation Guidelines

Our goal is to build a binary classifier able to automatically identify if sentences in financial annual reports are either *certain* or *uncertain*. For this purpose, we have to create a training set of human-annotated sentences based on which the classifier can learn. The following summarizes which categories of *uncertainty* we distinguish:

- Speculative statements about the future (e.g. “we view ... as a possibility in the future”)
- Believes or hopes (e.g. “we believe to be sufficiently prepared for ...”)
- Subjunctive phrases (e.g. “we might expand our portfolio ...”)
- Statements conveying uninformedness or doubt (e.g. “there is disagreement about ...”)
- Statements referring to economic uncertainty (e.g. “we view ... as an anomaly”)

However, this overview should only be taken as a general guideline—not every sentence containing a strong modal verb is automatically conveying uncertainty. The concept in its subtlety requires to pay close attention to the context in which such lexical markers are placed. We will further elaborate on this by means of examples. According to our definition, we would deem the following two sentences as *uncertain*:

1. “In addition, if we incur additional indebtedness to finance these acquisitions, the related risks we face from our already substantial level of indebtedness could intensify.”
2. “Such payment adjustments may adversely affect the results of our operations.”

In contrast, we would regard the following two as *certain*:

1. “Under the terms of the Plan, officers, directors, employees and other individuals may be granted options to purchase the Company’s common stock at exercise prices determined by the Company’s Board of Directors.”
2. “The price paid for fingerlings ranges between .75(cent) and 1.50(cent) per inch.”

The statement of the first sentence is not uncertain, despite the occurrence of the strong modal verb “may”. The second sentence contains a price range, which could be regarded as an imprecision in everyday language. Such ranges are common practice in financial disclosures, however, and should not bewilder a potential reader.

When labeling the sentences (“u” for uncertain, “c” for certain), also mark the words or phrases causing the sentence to be part of either category. This will help both you and us to understand what made you come to your conclusion. Lastly, avoid to continuously annotate for longer than one hour without a break. This will prevent you from the “fatigue effect” causing your annotation quality to suffer.