Neural models for Factual Inconsistency Classification with Explanations

Annotation Guidelines

The annotators used the following detailed annotation guidelines.

Overall aim: Given a pair of (claim, context) sentences, goal is to annotate the inconsistency to enable accurate localization of inconsistency in both claim and context. Further the goal is to also characterize type of inconsistency, and also categorize type of entity involved (both at a coarse and a fine-grained level).

Guidelines for Syntactic Oriented Annotations

Please keep in mind the following steps of the annotation process.

- Given a pair of (claim, context) sentences, read carefully and find out where the inconsistency lies, without using any other external reference or knowledge base.
- Highlight the "Source Chunk", "Relation Chunk" and "Target Chunk" in the claim, and the "Inconsistent Span" in the context, which are involved in the inconsistency as identified above.
  - "Source Chunk" is the linguistic chunk containing the entity lying to the left of the main verb/relating Chunk.
  - "Relation Chunk" is the linguistic chunk containing the verb/relation at the core of the identified inconsistency.
  - "Target Chunk" is the linguistic chunk containing the entity lying to the right of the main verb/relating chunk.
  - "Inconsistent Span" is the chunk in the context sentence that is inconsistent with the source, relation and target chunks identified in the claim.
- Further, for each of the source, relation and target chunks, identify the head and the modifier, and then label the inconsistent claim component as one of the Subject-Head, Subject-Modifier, Relation-Head, Relation-Modifier, Target-Head or Target-Modifier. Head is the linguistic head of the chunk and modifier is the remaining part of the chunk.

Detailed notes:

1. If you are unsure about the parts of the sentence use https://corenlp.run/ for the specific sub-phrase you cannot break.
2. Source or target chunks can be compound nouns, sometimes that will be evident through the context. We prefer the mismatch to be in the modifier as far as possible.

  - Consider this example where context is "I went through a railway tunnel." and claim is "I went through a big railway tunnel." Here "big" is a modifier of "railway tunnel" in the given claim.
  - Consider this example where context is "I went through a subway tunnel." and claim is "I went through a railway tunnel." Here "railway" is a modifier of "tunnel" in the given claim.
3. If a claim sentence has the form "⟨Source⟩ ⟨Relation⟩ ⟨Target1⟩ ⟨Target2⟩", then include only the relevant target in the span (according to where the mismatch is in the evidence). For example, consider claim as "Robert Downey Jr. starred in Inglourious Basterds and Sherlock Holmes.", and context sentence as "He was not involved in Inglourious Basterds.". Here "and Sherlock Holmes" will not be a part of any of the spans.
4. Finite verbs taking the main verb as a complement are part of the verb phrase (hence the relation): "We *tried to pass* OSN." Also, such finite verbs are relation modifiers.
5. When a sentence has only a subject and a verb in the present perfect tense ("X has Y'ed") then the auxiliary "has" is the relation and the past participle is the target.
6. For claims of the form "X ⟨verb⟩ ⟨preposition⟩ Y", if the verb is incomplete (semantically) without the preposition, it should be in the relation span; otherwise it should be in the target span with Y.
7. If the evidence does not contain an overt relation, but the inconsistency occurs in the relation of the claim, then in the evidence, mark the entity which is in focus and shared with the claim. For example, if claim is "Naruto is incapable of being a ninja." and context is "It tells the story of Naruto Uzumaki, an adolescent ninja who searches for recognition and dreams of becoming the Hokage , the leader of his village.", the inconsistent context span is "Naruto Uzumaki", although the inconsistent claim component is Relation-Modifier.

Guidelines for Semantic Oriented Annotations

Please keep in mind the following steps of the annotation process.

- You are given a claim sentence, context sentence, inconsistent claim triple (subject, relation, object) with head and modifiers, and inconsistent context span. Please do not use any other external reference or knowledge base.
- Label inconsistency type as one among these five types: simple, gradable, taxonomic relations, negation, set-based. Detailed type descriptions below.
- If an entity is involved in the inconsistency, label narrow entity type from among these 20 types: action, animal, entertainment, gender, geography, identity, material, name, nationality, organization, others, politics, profession, quantity, reality, relationship, sentiment, sport, technology and time.
- If an entity is involved in the inconsistency, label narrow entity type from among these 60 fine-grained types: action, animal, entertainment, entertainment-book, entertainment-brand, entertainment-category, entertainment-movie, entertainment-music, entertainment-role, entertainment-series, entertainment-technology,

gender, geography, geography-city, geography-continent, geography-country, geography-direction, geography-feature, geography-state, identity, identity-biology, identity-ethnic, identity-ideology, identity-social, material, name, name-actor, name-author, name-award, name-director, name-fiction, name-musician, name-university, nationality, organization, other, other-unknown, other-unrelated, politics, profession, profession-entertainment, profession-sport, quantity, quantity-cardinal, quantity-cardinal-money, quantity-ordinal, reality, relationship, sentiment, sport, sport-category, sport-outdoor, sport-outdoor-ball, technology, time, time-age, time-day, time-month, time-timeline, time-year.

Detailed description of inconsistency types:

- Simple: A simple contradiction is a direct contradiction, where the negative of one implies the positive of the other in a pair like *pass vs. fail*, i.e. X and not X. This also includes actions/ processes that can be reversed or have a reverse direction, like *come vs. go*, *ascend vs. descend* and *fill vs. empty*. Pairs with alternate viewpoints like *employer vs. employee*, *future vs. past* and *above vs. below* are also included in this category.
- Gradable: Gradable contradictions include adjectival and relative contradictions, where the positive of one, does not imply the negative of the other in a pair like *hot vs. cold*, *rich vs. poor least vs. most*, etc. This includes numbers (ordinals and cardinals), but not dates. This also includes periods of time (like three years or two minutes), but not points (like 2022 or yesterday).
- Taxonomic relations: We include three kinds of relations in this type: (a) Pairs at the same taxonomic level in the language like *red vs. blue* or *summer vs. spring* which are placed parallel to each other under the English color adjectives hierarchy. This also includes two taxonomically similar lists with a non-zero intersection. (b) When a pair has a more general word (*hypernym*) and another more specific word which includes the meaning of the first word in the pair (*hyponym*) like *giraffe* (hypo) vs. *animal* (hyper). hypo is a relation of inclusion, like: dog ⟨hypo to⟩ animal. hyper is the opposite direction. (c) Pairs with a part-whole relation like *nose vs. face* and *button vs shirt*.
- Negation: This includes inconsistencies arising out of presence of explicit negation particle in the sentence which is not an attached morpheme (e.g. *no*, *not*, *except*) or a finite verb negating an action (e.g. *fail to do X*, *incapable of X-ing*) etc. Also, "zero" is a negation morpheme. Tokens like "anything but", "anyone but," etc., are negation tokens and modifiers of the following entity.
- Set-based: This includes inconsistent examples where an object contrasts with a list that it is not a part of (e.g. *cat* vs. *dog, bee, and ant*).