XAlign: Cross-lingual Fact-to-Text Alignment and Generation for Low-Resource Languages

Tushar Abhishek, Shivprasad Sagare, Bhavyajeet Singh, Anubhav Sharma,
Manish Gupta, Vasudeva Varma
IIIT Hyderabad, India
{tushar.abhishek,shivprasad.sagare,bhavyajeet.singh,anubhav.sharma}@research.iiit.ac.in,{manish.gupta,vv}@iiit.ac.in

1 ETHICAL STATEMENT

The labeled dataset that we collected does not involve collection or storage of any personal identifiable information or offensive information at any stage. Human annotators were paid appropriately while performing data collection according to the standard wages set by National Translation Mission (https://www.ntm.org.in/) and mutually agreed upon. Although we have included the link to the dataset for reviewers' perusal, the data will be publicly released until MIT Open-Source License upon acceptance of this paper. The annotation exercise was approved by the Institutional Review Board of our institute.

2 MAIN TASK

The task is to mark English facts that are present in the given LR language sentence. You should choose all the facts that can be inferred from the given sentence by selecting the checkbox against it. Also, mention if the set of selected facts partially/completely cover the semantic information mentioned in the sentence.

3 INSTRUCTIONS RELATED TO PLATFORM

- When you select a question, you will see a sentence in low resource (LR) language and a list of English facts.
- Please read the LR sentence carefully. Although English translated sentence is provided for the reference, don't rely entirely on it. The translated sentence may not be accurate all the time.
- You will find list of English facts below the sentence. Please choose the facts that can be inferred from the given sentence by selecting the checkbox against it.
- If the sentence is grammatically incorrect, incomplete or erroneous for any other reason, please mention the reason in the textbox at the bottom.

4 INSTRUCTIONS RELATED TO ANNOTATIONS

- Exact Fact Matching: Information should exactly match what is present in the sentences (some exceptions are mentioned later; other than them, follow this rule strictly). For example,
  - Sentence: टीना मुनीम (जन्मः 11 फरवरी, 1955) हिन्दी फ़िल्मों की एक अभिनेत्री हैं।
  - English Translation: Tina Munim (DOB: 11 Feb 1955) is an actress who acts in Hindi movies.
  - Fact: Date of Birth | 11 February 1957.
  - Although the fact mentions that date of birth is 11 Feb 1957 but we won't consider it as a valid alignment for the sentence.
- Implied Information in facts
  - If information is related to language related inference and does not require external world knowledge (a piece of knowledge not embedded in language itself), we mark that fact.
    - Sentence: पी॰ नागराजन भारत की सोलहवीं लोकसभा में सांसद हैं ।
    - English Translation: P. Nagarajan is a Member of Parliament in India's 16th Lok Sabha.
    - Facts: P Nagarajan | position held | Member of the 16th Lok Sabha : P Nagarajan | occupation | politician.
    - For the given sentence, the information that the subject is a politician (राजनेता) isn't written, but we can say that a Member of Parliament will be a politician, hence we mark it.
    - As another example, consider a sentence that says that a person did her Masters in Geography but doesn't explicitly mention her occupation directly. Still, we can mark the occupation= geographer fact as valid.
  - If information in the fact requires external world knowledge, we do not mark that fact.
    - Sentence: अमृता मलयाली माँ और पंजाबी पिता की संतान हैं और वह मुंबई में पैदा हुई थी।
    - English Translation: Amruta's mother is a Malayali and her father is a Punjabi, and she was born in Mumbai.
    - Fact: Place of Birth | Chembur.
    - Even if you know that Chembur is in Mumbai, please don't mark it.
- If some facts contain redundant information , then dont mark it.
- Abbreviations: If the part of the sentence is abbreviated in the facts or if the part of fact is abbreviated in the sentence, we don't consider those facts.
  - Sentence: फील्ड मार्शल आर्किबाल्ड पेर्सियल वेवेल , पहले अर्ल वावेल , जीसीबी , जीसीएसआई , जीसीआईई , सीएमजी , वीएम , केएसटीजे , पीसी (5 मई 1883 – 24 मई 1950)

, ब्रिटिश सेना के एक वरिष्ठ अधिकारी और भारत के वाइसराय थे ।

-- English Translation: Field Marshal Archibald Percival Wavell, 1st Earl Wavell, GCB, GCSI, GCIE, CMG, MC, KStJ, PC (5 May 1883 – 24 May 1950) was a senior officer of the British Army and an Indian Viceroy.

-- Facts: Archibald Wavell, 1st Earl Wavell | award received | Virtuti Militari

- Fact Generalisation
  - If specific information is present in the sentence but there isn't an exact match in the fact list, then select the apt synonyms.
    - ∗ Sentence: उन्होने अपनी कविताओं से एक अच्छे साहित्यकार की छवि स्तापित कर ली थी
    - ∗ English Translation: He had established the image of a good litterateur through his poems.
    - ∗ Now if the fact list contains occupation as poet, and there is no other fact with occupation as litterateur, we consider the apt synonym and mark this fact as valid.
  - If facts contain more specific terms as compared to the term present in the sentence then consider that fact for annotation (facts can contain more specific information).
    - ∗ Sentence: राजगोपाल चिदम्बरम (जन्म 12 नवम्बर 1936) जिन्हें सामान्यतः आर० चिदम्बरम के नाम से जाना जाता है , पद्मविभूषण सम्मानित भारतीय वैज्ञानिक हैं ।
    - ∗ English Translation: Rajagopal Chidambaram (born 12 November 1936), commonly known as R. Chidambaram, is a Padma Vibhushan honored Indian scientist.
    - ∗ Fact: Rajagopala Chidambaram | occupation | nuclear physicist
    - ∗ We mark this fact as a nuclear physicist is also a वैज्ञानिक (scientist). The fact has more specific information and we mark it as valid.