Türkçe CoLA veriseti direktifleri

Orjinal CoLA (The Corpus of Linguistic Acceptability) 10.000 civarı cümleden oluşan bir veriseti, bu veriseti dildeki gramer açısından kabul edilebilir olan ve olmayan cümleleri içeriyor, sonuç olarak da dil modellerinin gramer bilgisini ölçmek için kullanılıyor.

Verisetindeki cümlelerse dilbilgisi kitaplarından toplanmış ve elle anotation yapılmış. Her cümle olduğu gibi, orijinal olarak (kabul edilebilir) veya değişik şekillerde değiştirilerek

(kabul edilemez) olarak verisetinde yer alıyor.

Benim verisetim TrCoLA bu setin fikir ve uygulama olarak aynısı, dilbilgisi kitaplarından toplamda 3900 civarı cümle topladım, kısadan uzuna sıraladım. Kabul edilemez varyasyonları ChatGPT ile ürettim, ChatGPTye cümleyi ve istediğim varyasyon çeşidini sundum. Bu örnekte olduğu gibi:

Horul horul uyudunuz. Orijinal

Horullar horul uyudunuz. Morphological variation

Horul uyudun horul. Syntactic variation

Horul horul ağladınız. Semantic variation

Örnekte de gördüğünüz gibi, yanlarındaki üretme biçimine göre orijinal cümlenin değiştirilmesiyle oluşuyorlar ve variationların hiçbiri gramere uygun cümleler değiller.

Variation Tipleri

Verisetinde toplam 3 variation var, morfolojik, syntactic ve semantic. Morfolojik variation tipi bir veya birden fazla sözcüğün biçim ollarak bozulmasıyla yapılıyor, sözcük tipine uygun olmayan ekler eklemek veya ekleri yanlış şekilde eklemek gibi:

Gittim gittim**ken**

Gittim gittim**e**

Evde ev**se**de

Evde evde**den**

Değişiklikler eklerde olmalı, köklerde olmamalı.

Syntactic variation ise cümle sözdizimi bozularak yapılıyor, bu çeşitte sözcüklerin yerleri değişebilir veya sözdiziminii bozacak bazı sözcükler eklenebilir:

Pek yorgun görünüyordu . Pek yorgun değil görünüyordu.

Gün daha doğmamıştır. Gün daha doğmamıştır. Mı?

Akşamları kitap okur. Kitap akşamları okur.

Son tipte ise cümleyi anlamsal ollarak bozmak amaçlanır. Bu tip genel olarak cümleye anlamsal olarak uymayan sözcüklerin eklenmesi veya bazı sözcüklerin çıkarılarak/değiştirilerek cümle anlamının bozulmasıyla olur.

Öznesiz nesne olmaz. Öznesiz nesne var.

Sonbahar yaklaşmakta. Gökyüzünde balık uçmakta.

Bu tipte sözdizimsel veya biçimsel olarak bir bozukluk yoktur, ancak cümleler anlamlı değidir. Bazı gerçekleri anlatan cümlelerdeki gerçeklerin tersine çevrilmesi veya gerçeğin değiştirilmesi bu türe girmez, sonuç varyasyon net şekilde anlamsız olmalıdır:

Marmara Denizindeki balıklar max 5 kg'a ulaşır. orijinal

Marmara Denizindeki balıklar max 5 metre uçar. EVET, bu bir semantic varyasyon çünkü anlamsız

Marmara Denizindeki balıklar max 3 kg'a ulaşır. HAYIR, cümle anlamlı o yüzden bu bir varyasyon değil

Veriseti Formu

Önce bir örnek:

{

"id": "s46",

"orig": "Kimseyi görmedim.",

"variation": "Kimseysi görmedim.",

"var_type": "Morphological Violation"}

Örnekten de açık olduğu üzere her örnek bir id, orijinal cümle, üretilmiş varyasyon ve varyasyon tipinden oluşuyor.

Task

Başta yazdığım gibi varyasyonarı ChatGPT ile ürettim, biindiği üzere ChatGPT benzeri modeer her zaman doğru sonuçları üretmiyor. Yapılmasını istediğim task, doğru olmayan örneklerin verisetinden çıkarılması. Her şeyden önce tek bir kural var, **olabildiğince az örnek atmak.**

Öncelikle yukarda verilen 3 tipi ayırt etmenin pratik yollarını görelim:

1. Morfolojik varyasyonları görmek daha kolay, cümledeki sözcük kökeri aynıysa ve bazı sözcüklerin ekerinde değişikler varsa ve bu değişiklikller sonucu cümle gramer kurallarına uygun bir cümle olmaktan çıkıyorsa o zaman bu cümle bu tipe aittir.
2. Syntactic variationda sözcük değişimleri olabilir veya olmayabilir. Sözcük değişimi yoksa sözcüklerin yerleri değişmiştir, böylece cümle gramer-dışı bir şekle girmiştir ve anlamını yitirmiştir. Sözcük değişimleri de sözcük ekleme veya çıkarma şeklinde olur, eklenen sözcükler cümlenin asıl anlamını modifiye etmez ama sözdizimibin bozar.
3. Semantic variationda genelde sözcük değişimeri olur ve cümle hepten anlamsız hale gelir.

2 ve 3 ü karşılaştıralım:

Mavi balinalar 200 tona ulaşabilir. Orijinal

Mavı balinalar 200 tona ki ulaşabilir. Syntatic varyasyon

Mavi balinalar 200 ton uçabilir. Semantic variation

Artık kurallara bakabiliriz:

Kural 0: Varyasyon anlamlı ve kurallı bir cümleyse, okuyunca ne anlamsal ne de başka şekilde gözünüze yanlış gelmiyorsa bu örneği verisetinden çıkarmalısınız. Unutmayın, bi anlamı bozulmuş varyasyonları arıyoruz.

Kural 1: 2 ve 3 arasında kaldıysanız ve bir şekilde ayırt edemiyorsanız, örnek olduğu gibi kalabilir. Önemli olan varyasyonun bu 3 tipten birine uyacak şekilde bozuk olması, varyasyon tipinden %100 emin değiseniz ve arada kaldıysanız örnek verisetinde kalmalı.

Kural 2: Varyasyon anlamsız bir cümleyse, bu 3 tipten birine uygunsa ama varyasyon tipi gözünüze yanlış geliyorsa bu örnek verisetinde kalmalı. Unutmayın biz bütün varyasyonları arıyoruz, tipin yanlış olması tabi ki iyi değil ama varyasyon cümlenin kendisinden daha önemli değil. Örnek:

Ben de gittim. - Ben de gittimken Syntacti variation (yanlış, doğrusu morfolojik olacak) Bu örnek sette KALMALI

Kural 3: Orijinal cümle ve varyasyon benziyorsa ve varyasyon anlamsızsa ancak varyasyon tipi bu 3 örnekten hiçbirine uymuyorsa, örneği çıkarmalısınız.

Kural 4: Orijinal cümle ve varyasyon birbirine benzemiyorsa, ancak varyasyon başka bir cümlenin bu 3 tipten birine uygun varyasyonuysa bu cümle kalmalı. Örnek:

Balinalar uçamaz. Köpekler tek oturuşta 200 kg pencere yiyebilir. - Semantic Varyasyon Bu örnek KALMALI, çünkü 2 numaralı cümle köpeklerle ilgili bir cümlenin varyasyonu olabilir.

Özet olarak:

- Varyasyon her zaman anlamsız olmalı.
- Varyasyon tipi her zaman bu 3 tipten biri olmalı.
- Tutabildiğimiz kadar çok örneği tutmalıyız, amaçlardan biri olabildiğince büyük ve doğru bir set çıkarmak.

Veriseti

Veriseti: tek dosya, **tr_cola_variations.jsonl**

Büyüklük: 10851 satır