Instrukcja Anotacji

Polish Question Answering Dataset

20 maja 2022 r.

Wprowadzenie

Zadanie anotacyjne polega na przygotowaniu pytań do tekstów, i zaznaczeniu w tekście odpowiedzi, oraz podania formy bazowej odpowiedzi. Każdy tekst składa się z dwóch fragmentów artykułu z polskiej wikipedii, 1. streszczenia (całego tekstu artykułu znajdującego się przed pierwszym nagłówkiem) 2. wybranego paragrafu ze środka artykułu. Anotacji podlega tylko fragment nr 2, streszczenie służy do wprowadzenia anotatora w temat, na wypadek niezrozumiałości fragmentu nr 2 w oderwaniu od kontekstu całego artykułu. Paragrafy do anotacji zostały wybrane na podstawie automatycznych metod oceny istotności paragrafu dla całego artykułu, oraz długości paragrafu. Metody te pozwalają szacować wartość paragrafów do anotacji, natomiast nie są idealne, i część paragrafów (oraz artykułów) może okazać się nieinteresująca, albo nie pozwalać na przygotowanie wartościowych pytań. Takie paragrafy należy pomijać (lub zadawać do nich tylko tyle pytań, na ile pozwalają), pytania wymyślane na siłę, niskiej jakości, mogą bardziej zaszkodzić niż pomóc.

Priorytetem powinna być jakość anotacji. Pojęcie to jest precyzowane przez resztę tej instrukcji. Niektóre z wymienionych dyrektyw mogą się wydać sztuczne, lub nieelastyczne, natomiast stosowanie ich jest istotne, aby zapewnić spójność danych.

Instrukcja zadania

> *Poświęć około 8 minut na przygotowanie dla wskazanego tekstu 5 pytań. Wśród nich powinno się znaleźć jedno pytanie "nieodpowiadalne". Dla każdego pytania, zaznacz minimalny fragment z podanego paragrafu, zawierający na nie odpowiedź, dla pytań nieodpowiadalnych zaznacz minimalny fragment który pozornie zawiera odpowiedź. Dla każdej odpowiedzi która różni się od swojej formy bazowej (czyli takiej, którą podalibyśmy odpowiadając na pytanie w sposób naturalny), podaj jej formę po normalizacji. Jeżeli nie możesz zadać więcej pytań do podanego tekstu, zaznacz to w odpowiednich polach. Unikaj używania fraz i słów pojawiających się w kontekście odpowiedzi. Zachęcamy do zadawania pytań trudnych.*

Typy pytań

Typ odpowiedzi

Ze względu na typ odpowiedzi (a zatem i sposób anotacji) możemy wyróżnić dwa zasadnicze typy pytań:

1. pytania o rozstrzygnięcie
2. pytania o dopełnienie

W wypadku pytań o rozstrzygnięcie, naturalna odpowiedź na takie pytanie musi móc brzmieć “Tak” lub “Nie”. Przykłady:

> Czy Lenin zmarł śmiercią naturalną?

> Czy psy są wszystkożercami?
>
> Czy encefalopatia gąbczasta jest chorobą śmiertelną?

Do tej kategorii nie wpadają jednak pytania o wybór spośród kilku opcji:

> Czy pierwsza wojna światowa skończyła się przed, czy po przejęciu władzy w Rosji przez bolszewików?
>
> Czy ziemniak to owoc, czy warzywo?
>
> Który naukowiec jako pierwszy uzyskał czysty tlen: Lavoisier, Scheele czy Priestley?

Również tam, gdzie opcje są *de facto* dwie, natomiast nie są wymieniane explicite:

> Jaka jest ocena polityki interwencjonistycznej w monetaryzmie?

Pytania o dopełnienie możemy rozumieć jako wszystkie pytania, które nie są pytaniami o rozstrzygnięcie (a więc wpadają tu także pytania które wykluczyliśmy z pytań o rozstrzygnięcie powyżej). Przez ten negatywny charakter definicji, trafiają tu pytania bardzo różnorodne. Można próbować wskazać główne kategorie takich pytań, ale takie wyliczenie nie ma charakteru skończonego, i nie powinno ograniczać naszej kreatywności przy wymyślaniu pytań:

1. Pytania o miejsce - *Gdzie leżą szczątki Chopina?*
2. Pytania o czas - *Kiedy zmarł Chopin?*
3. Pytania o liczbę - *Ile lat miał Chopin w momencie śmierci?*
4. Pytania o osobę - *Kto był ojcem Chopina?*
5. Pytania o przyczynę - *Z jakiego powodu Chopin emigrował do Francji?*
6. Pytania o skutek - *Do czego doprowadziło wyrzucenie fortepianu przez okno?*
7. Pytania o cel - *W jakim celu Chopin napisał Etiudę Rewolucyjną?*
8. Pytania o wyliczenie - *Którzy kompozytorzy stanowili główne inspiracje artysty?*

   …

Odpowiadalność

Dodatkowo wyróżniamy klasyfikację ze względu na ich relację do tekstu do którego są przyporządkowane:

1. Pytania odpowiadalne
2. Pytania nieodpowiadalne

Pytania odpowiadalne to takie, co do których w tekście znajduje się odpowiedź. Obecność odpowiedzi jest kwestią płynną, na jednym krańcu spektrum, mamy literalne znajdowanie się

w tekście, na drugim kompletną niezależność logiczną od tekstu. Weźmy pod uwagę następujący fragment:

> *Praca pedagogiczna na uczelni bardzo go absorbowała – zarzucił ją dopiero na trzy lata przed śmiercią. Pomimo tego znajdował czas na prowadzenie samodzielnych badań filozoficznych. Spędził na nich następną dekadę, czego efektem była publikacja napisanej w scholastycznym języku obszernej Krytyki czystego rozumu w 1781 roku – jednego spośród ważniejszych dzieł w historii filozofii. Ponieważ ta praca spotkała się z małym odzewem, w 1783 roku Kant wydał skromniejsze objętościowo i bardziej przystępne Prolegomena, zawierające wykład jego głównych idei. Pozostałe publikacje Kanta z okresu krytycznego to Uzasadnienie metafizyki moralności z 1785 roku, będące uproszczoną wersją Krytyki praktycznego rozumu z 1788 roku, oraz Krytyka władzy sądzenia z 1790 roku. W swych dziełach zajął się kolejno teorią poznania, etyką oraz estetyką. Pod koniec tego okresu pozostawał Kant pod wpływem empiryzmu Hume'a.*

Oraz zadane do niego pytania, układające się zgodnie z przebiegiem spektrum odpowiedź obecna w tekście - odpowiedź nieobecna w tekście:

1. W którym roku ukazała się Krytyka Czystego Rozumu?
2. Ile lat trwały badania prowadzące do publikacji z 1781?
3. Jakimi działami filozofii zajmował się Kant pod koniec XVIII wieku?
4. Dlaczego Kant porzucił plany wydania Krytyki Czystego rozumu?
5. W którym roku ukazała się najważniejsza książka królewieckiego filozofa dotycząca deontologii?
6. Jakie jest miejsce urodzenia Hume'a?

1. Odwołuje się do daty która jest literalnie wymieniana w tekście, w zdaniu mówiącym o ukazaniu się Krytyki Czystego Rozumu. Odpowiedź na pytanie w tekście jest tak bezpośrednia, że pytanie robi się banalne.

2. Również znajduje odpowiedź na pytanie w tekście, natomiast wymaga ona pewnego rozumowania o charakterze lingwistycznym (znajomości znaczenia słowa "dekada"), to czyni owo pytanie bardziej wymagającym.

3. Wymaga kompetencji lingwistycznych (umiejętności rozpoznania, że "swych" odnosi się do Kanta) dodatkowo wiedzy o świecie, natomiast jest to wiedza, której można się spodziewać od przeciętnego człowieka. Wymagamy tu bowiem umiejętności rozpoznania, że lata 1780-te i 1790-te wpadają do końca XVIII wieku, oraz, że teoria poznania, etyka, i estetyka to działy szerszej dziedziny jaką jest filozofia.

4. Jest poprawnym pytaniem nieodpowiadalnym, dotyczy Kanta (tematu artykułu), dotyczy Krytyki Czystego Rozumu (ważnego elementu paragrafu), natomiast tworzy fałszywe połączenie w tekście. Wspomniana książka ukazała się drukiem, więc pytanie zawiera nieprawdziwe założenie, które dodatkowo łączy z innym fragmentem w tekście w sposób, który nie jest uzasadniony, natomiast wykorzystuje pewne fakty na temat tekstu, aby stworzyć pozory poprawności. Fakty te to, możliwy (ale nieprawdziwy) związek między absorbującą pracą na uczelni i zaniechaniem innych obowiązków, występowanie w zdaniu zawierającym odpowiedź pozorną zbliżonych do pytania wyrazów ("zarzucił"), oraz mowa o Krytyce Czystego Rozumu, i jej problemach w tekście (słaby odzew). Pytania nieodpowiadalne nie muszą, ale mogą opierać się o fałszywe założenia.

5. Jest pytaniem nieodpowiadalnym. Zakłada bowiem dość specjalistyczną wiedzę zarówno lingwistyczną (“królewiecki filozof” jako alias Kanta), jak i z dziedziny filozofii (że Krytyka Praktycznego Rozumu jest głównym wykładem deontologii - kantowskiej teorii moralnej). W tym wypadku zaznaczona odpowiedź jest poprawna (oraz pytanie nie opiera się na błędnych założeniach), ale uznajemy pytanie za nieodpowiadalne, ponieważ w tekście nie ma dostatecznej ilości informacji, żeby człowiek z ulicy wyciągnął z niego takie wnioski. Nie oznacza to, że jest to **dobre** pytanie nieodpowiadalne, zauważmy że w tekście w ogóle nie pojawia się słowo “deontologia”, dodatkowo wskazana data nie sprawia wrażenia lepszej odpowiedzi, niż jakakolwiek inna pojawiająca się w tekście (w tekście nie ma nic, co skłaniałoby nas do hipotezy, że to 1788 a nie np. 1790 lub 1785 jest poprawną odpowiedzią). Takie pytanie więc odrzucamy, jako nie posiadające dostatecznie sugestywnej odpowiedzi pozornej.

6. Jest pytaniem nierelewantnym (a więc nieodpowiadalnym) do tekstu. Owszem, tekst wspomina o Hume’ie, natomiast nie mówi nic o jego miejscu narodzenia, ani nawet nie zawiera żadnej sugestii, pozwalającej formułować jakieś hipotezy. Oczywiście pytania mogą być nierelewantne względem tekstu w bardziej ekstremalny sposób (np. “Z czego otrzymuje się podpuszczkę?”). Pytania nierelewantne nie są dla nas interesujące.

Z całego tego spektrum interesują nas więc pytania od 2. do 4. - 1. jest raczej zbyt banalne, 5. raczej zbyt trudne.

Nieodpowiadalność a prawdziwość

Nieodpowiadalność nie jest relacją między pytaniem a światem, tylko między **pytaniem** a **tekstem**. Należy zwrócić uwagę na to, że takie określenie tej relacji (relacja między pytaniem a tekstem), implikuje, że pytanie jest odpowiadalne albo nieodpowiadalne niezależnie od tego, jaką odpowiedź zakreślimy w tekście. Zaznaczywszy pozorną odpowiedź, to czy jest ona prawdziwą czy fałszywą odpowiedzią na pytanie nie powinno nas interesować, interesuje nas tylko to, czy na podstawie tekstu (zakładając jego prawdziwość) możemy stwierdzić, że dany fragment odpowiada na pytanie. Z drugiej strony jeśli w tekście znajduje się odpowiedź na pytanie, ale dodatkowo możemy w nim znaleźć odpowiedź pozorną (i to ją zakreślimy), w takiej sytuacji pytanie to **nie** jest nieodpowiadalne.

Dobre pytanie nieodpowiadalne jest więc 1. nieodpowiadalne (nie ma w tekście fragmentu stanowiącego według tekstu odpowiedź na to pytanie), 2. ma pozorną odpowiedź (jest w tekście fragment, który dla nieuważnego czytelnika mógłby sprawiać wrażenie odpowiedzi). W takiej sytuacji jako odpowiedź na pytanie nieodpowiadalne zakreślamy fragment wyrażający odpowiedź pozorną, i dodatkowo oznaczamy pole “Pytanie nieodpowiadalne” ptaszkiem.

Typy nieodpowiadalności

Podobnie jak w wypadku typów pytań o dopełnienie, możemy wyróżnić kilka podstawowych typów nieodpowiadalności, w zależności od tego, dlaczego dane pytanie jest nieodpowiadalne. Ponownie, klasyfikacja ta nie jest wyczerpująca, i nie ma za zadanie ograniczania naszej kreatywności, a raczej jej pobudzenie, przez wskazanie przykładów.

> *Pierwsze wersje granatów składały się ze skorupy z papieru, ceramiki lub prymitywnego szkła, wypełnionej prochem i czasem dodatkowo siekańcami*

metalowymi lub substancjami mającymi zwiększyć ich działanie bojowe, np. kwasem, substancjami zapalnymi, lub drażniącymi (np. wapno palone). Zapalnik miał postać lontu, który należało podpalić przed rzuceniem, co utrudniało ich użycie. Granaty tej postaci pojawiły się w Chinach za czasów dynastii Tang. Jednym z najstarszych znanych wyobrażeń jest malowidło ścienne z grot Mogao w Dunhuangu. Na większą skalę granaty rozpowszechniły się na przełomie wieków XVI i XVII.

1. Wzmocnienie tezy w tekście:
   Jakie jest **pierwsze** w historii świadectwo użytkowania granatów?

2. Fałszywe połączenie:
   Z jakiego materiału jest wykonane malowidło przedstawiające granaty w grotach Mogao?

3. Fałszywe założenie:
   Dlaczego granaty zniknęły z pól walki?

4. Odwrócenie relacji w tekście:
   W którym wieku granaty przestały być popularne?

1. Jest dość prostym sposobem tworzenia pytań nieodpowiadalnych, polega na wzmocnieniu informacji zawartej w tekście, np. w tekście jest mowa o tym, że większość X jest Y w warunkach Z, pytamy “W jakich warunkach wszystkie X są Y?”. Łatwo rozpoznać takie pytania poprzez stosowanie kwantyfikatorów ogólnych takich jak “zawsze”, “wszystkie”, “pierwszy”, “ostatni”, “najwyższy” itd. są one stosunkowo nienaturalne, i często wyglądają podejrzanie (i.e. domyślamy się, że są podchwytliwe) dlatego nie należy ich nadużywać.

Proporcje pytań

Dane powinny być możliwie różnorodne. Pewne typy pytań (pytania o daty, nazwiska, nazwy miejsc) są prostsze do tworzenia niż inne, nie oznacza to, że są one niedozwolone, natomiast nie mogą one naruszać właściwych proporcji. W szczególności w zbiorze danych ok 20% pytań, powinno być pytaniami nieodpowiadalnymi, co przekłada się na przybliżony stosunek 4:1 per paragraf. Stosunek ten może być różny w różnych paragrafach - w szczególności dopuszczalne jest zadanie np. tylko jednego pytania odpowiadalnego, i aż trzech pytań nieodpowiadalnych, jeżeli jest to kompensowane w innych paragrafach.

Należy przede wszystkim zwrócić uwagę na niezależność stochastyczną cechy odpowiadalności, od cechy bycia pytaniem o rozstrzygnięcie. To znaczy że proporcja pytań o rozstrzygnięcie wśród pytań nieodpowiadalnych (np. 21%) powinna być podobna do proporcji pytań o rozstrzygnięcie wśród pytań odpowiadalnych (np. 18%). Zaburzenie takiej proporcji jest poważną wadą anotacyjną.

Jeżeli chodzi o typy pytań ze względu na typ odpowiedzi, możemy orientacyjnie przyjąć (bardzo zgrubnie) następujące proporcje docelowe:

1. Pytania o miejsce ~ 6.5%

2. Pytania o czas ~ 6.5%
3. Pytania o osobę ~ 6.5 %
4. Pytania o liczbę ~ 6.5 %
5. Pytania o przyczynę/skutek/cel/rację/dowód ~ 20 %
6. Pytania o wyliczenie ~ 10 %
7. Pytania o rozstrzygnięcie ~ 15 %
8. Pozostałe ~ 30 %

Pierwsze cztery kategorie otrzymują tak niskie proporcje z tego względu, że są bardzo proste do rozpoznania (z wykorzystaniem słów pytajnych “Gdzie”/”Skąd”, “Kiedy”, “Kto”/”Kogo”, “Ile”) i odpowiedzenia (daty, liczby, nazwiska, nazwy miejsc itd. są stosunkowo proste do wykrycia w tekście). Model uczony tylko takich zadań, będzie skłonny stosować mało wyrafinowane sposoby przeszukiwania tekstu.

Niejednoznaczność pytań

Znaczna część naszego języka jest niejednoznaczna, i dotyczy to również pytań. Również pod względem niejednoznaczności, możemy uszeregować pytania w spektrum. Na przykład:

1. W którym roku Henryk VIII po raz pierwszy się ożenił?
2. Kiedy Henryk VIII po raz pierwszy się ożenił?
3. Kiedy Henryk VIII się ożenił?
4. Kiedy król się ożenił?
5. Kiedy on się ożenił?
6. Kiedy to się stało?

Nie chcemy od tego zjawiska uciekać, a więc akceptujemy pewną dozę niejednoznaczności. Natomiast chcemy unikać 1. bardzo wysokiego stopnia niejednoznaczności (pytań zbliżonych do 5. i 6.), 2. powszechnej niejednoznaczności (czyli sytuacji w której w każdym pytaniu w zbiorze danych jest wyraźny stopień niejednoznaczności - ponownie, zachowujmy proporcje).

Możemy wyróżnić kilka źródeł niejednoznaczności:

Zaimki

> *Po przegranym starciu pod dowództwem Jana Hunyadyego **Murad II** został zmuszony do abdykacji na rzecz syna. Powrócił na tron w 1446 (panował do 1451).*

Jakie wydarzenie doprowadziło do zrzeczenia się przez **niego** tronu?

> *W 1282 roku władze w **Austrii** objęli Habsburgowie, którzy w 1453 roku przyjęli tytuł arcyksiążąt.*

Która dynastia wstąpiła **tam** na tron w XIII wieku?

W obu wyżej wymienionych przykładach pojawiają się zaimki, które w oderwaniu od zestawionego z pytaniami artykułu, stają się niezrozumiałe - nie wiemy o czyją abdykację

pytamy, bez znajomości kontekstu - że artykuł dotyczy biografii Murada II. Jeżeli paragraf jest poświęcony jednej osobie/miejscu/zwierzęciu/pierwiastkowi chemicznemu/rodzajowi sera itd. uznajemy, że jest jasne do czego mogą się odnosić zaimki (do tej rzeczy). Analogicznie jest w sytuacji, kiedy artykuł dotyczy czegoś innego, ale cały paragraf jest wyjątkowo poświęcony rzeczy o którą pytamy w pytaniu. Niejednoznaczności wynikającej z zastosowania zaimków nie dopuszczamy w sytuacji w której w danym paragrafie jest kilka obiektów co do których można mniemać, że zaimek się odnosi, np.

> *Szybko stał się bastionem obrony ortodoksji chrześcijańskiej, przeciwstawiając się doktrynom monofizyckim zawartym w Henotikonie cesarza Zenona Izauryjczyka. Klasztor nie przystąpił też do schizmy patriarchy Konstantynopola Akacjusza (484–519). W VIII-IX w. klasztor mocno był zaangażowany w walkę z ikonoklazmem. Igumen Saba wykazał nieprzejednane stanowisko w tej sprawie na soborze nicejskim II w 787 za cenę konfliktu z patriarchą Metodym Wyznawcą. Pod rządami igumena św. Teodora Studyty (zm. 826) klasztor stał się modelem monastycyzmu wschodniego i wiodącym przeciwnikiem ikonoklastów Nicefora I Genika i Leona V Armeńczyka.*

W których latach sprawował **on** urząd [patriarcha Akacjusz]?

Zwróćmy uwagę, że w tekście jest wymieniony tylko jeden okres urzędowania (reszta to pojedyncze daty). Natomiast w tekście w sposób równorzędny pojawia się wielu patriarchów, nie możemy więc mieć pewności, że to o Akacjusza chodzi w pytaniu.

Czasowniki w formie osobowej

Tutaj sytuacja jest bardzo podobna, czasem w języku polskim stosujemy podmiot domyślny, i wówczas na podstawie formy czasownika i kontekstu możemy się domyślać o kogo chodzi.

> *Początkowo uczyła się w domu, później **Hepburn** zaczęła uczęszczać do Bryn Mawr College.*
>
> Na której uczelni **studiowała**?

Analogicznie: jeżeli podmiotem domyślnym czasownika z pytania jest obiekt będący głównym tematem artykułu lub paragrafu, lub nie ma w tekście wyraźnych konkurentów, taka niejednoznaczność jest dopuszczalna.

Odniesienie rzeczowników pospolitych

Odniesienie rzeczowników pospolitych również może być niejednoznaczne np.

> *17 maja została odprawiona msza, celebrowana przez kardynała Aleksandra Kakowskiego.*
>
> Kiedy miała miejsce msza?

Rzeczownik “msza” w tym wypadku odnosi się do mszy żałobnej z okazji pogrzebu Józefa Piłsudskiego. Bez znajomości kontekstu, nie wiemy jednak o którą mszę chodzi, więc pytanie staje się niejasne. Jeżeli takie pytanie pojawiłoby się w artykule, lub paragrafie dedykowanym pogrzebowi Piłsudskiego, można by zakładać jednoznaczność na podstawie kontekstu, natomiast w sytuacji gdy msza odgrywa w paragrafie rolę drugorzędną, powinno to być odzwierciedlone w sposobie zadania pytania.

Sposoby redukcji niejednoznaczności

Najprostszym sposobem redukcji niejednoznaczności, jest dookreślenie poprez dodanie okoliczników. W ten sposób możemy np. doprecyzować powyższe pytanie do postaci:

> Kiedy miała miejsce msza żałobna z okazji śmierci Marszałka?

Pokrycie leksykalne

Istotną cechą pytań, układanych do tekstów, jest stosunek wspólnych słów pomiędzy pytaniem, i zdaniem (lub zdaniami) zawierającym(i) odpowiedź. Pokrycie leksykalne, to stosunek liczby słów pojawiających się w pytaniu i zdaniu zawierającym odpowiedź, do liczby słów w pytaniu. Przy liczeniu pokrycia leksykalnego pomijamy różnice fleksyjne, oraz nie liczymy słów funkcyjnych (takich jak “w”, “swój”, “ona”, “ale”, “gdy”).
Na przykład:

> *30 maja 1966 wyszła za mąż za Carla Thomasa Deana, którego poznała dwa lata wcześniej w pralni w Nashville.*

1. Gdzie Dolly Parton **poznała** swojego **męża**? 2/4
2. Gdzie Dolly Parton pierwszy raz spotkała swojego późniejszego małżonka? 0/7

1. dzieli ze zdaniem zawierającym odpowiedź na pytanie dwa słowa z pięciu (pomijamy “swojego”, i “Gdzie”), czyli 50% całego pytania. W drugim pytaniu, nie mamy żadnego pokrycia, co czyni to pytanie trudniejszym (wymaga więcej kompetencji językowych) a więc i bardziej wartościowym.

Oczywiście w wypadku niektórych pytań (np. takich, które zawierają nazwy własne, np. tytuły filmów), ciężko jest uniknąć pewnego pokrycia.

> *Z kolei ballada „I Will Always Love You” (1973) nagrana przez Whitney Houston na potrzeby filmu Bodyguard (1992) okazała się megahitem na całym świecie.*
>
> Aranżacja której artystyki rozsławiła utwór “**I Will Always Love You**”? 5/9

W takich sytuacjach nie należy unikać powtarzania słów składających się na dany tytuł, może to prowadzić do sztuczności i niejasności.

W pozostałych przypadkach, możemy wykorzystać szereg środków, aby zredukować pokrycie:

Początkowo Leeuwenhoek zajmował się kupiectwem (prowadził sklep z galanterią męską), a nocami szlifowaniem szkieł oraz konstrukcją mikroskopów.

1. Jaki sklep prowadził Leeuwenhoek? 3/3
2. W jakiej branży handlował? 0/2
3. W jakiej branży operowało jego przedsięwzięcie handlowe? 0/4

Pomijamy pojawiające się w tekście nazwisko, możemy to zrobić dzięki formie osobowej czasownika, w innej sytuacji moglibyśmy zastąpić nazwisko zaimkiem, lub rzeczownikiem pospolitym (“Badacz”, “Wynalazca”, “Naukowiec”). Zamiast “prowadził sklep”, w 2. skracamy to wyrażenie do jednego słowa o podobnym znaczeniu (“handlował”), w drugą stronę możemy wydłużyć jedno ze słów (“sklep”), do bardziej rozbudowanego wyrażenia (“przedsięwzięcie handlowe”) w 3. Podobną techniką jest stosowanie wyrażeń o odmiennej precyzji: ogólniejszych, lub o węższym zakresie.

Próg pokrycia leksykalnego, powyżej którego potrzebna jest ręczna akceptacja anotacji wynosi 50%.

Ciekawość

Pytania powinny być ciekawe. Najlepiej wyjaśnić to pojęcie kontrastując je z przeciwnym krańcem spektrum, gdzie mamy pytania mało interesujące, które możemy podzielić na kilka głównych kategorii:

Pytania nadmiernie zależne od struktury tekstu

Nieinteresujące są pytania które zamiast pytać o sam temat tekstu, są mocno zapośredniczone w sposobie opisu tematu przez autora tekstu. Opisując jakiś temat, autor podejmuje szereg decyzji o charakterze kompozycyjnym nt. tego, co jest istotne, a co nie. W szczególności za istotny uznaje sam temat tekstu. Decyzje te nie muszą odzwierciedlać faktycznych, albo uznanych konwencjonalnie proporcji nt. istotności tych rzeczy. Problem ten dotyczyć będzie więc przede wszystkim sytuacji, w których mamy do czynienia z wielością jakichś elementów, natomiast tekst akcentuje tylko jeden z nich.

Główne grupy które możemy wyszczególnić:

1. Wielość przyczyn, celów, zastosowań

Hipnozę wykorzystuje się w kryminalistyce, w procesie identyfikacji sprawców.

Gdzie stosuje się hipnozę?

Dostaliśmy paragraf który mówi o zastosowaniach kryminalistycznych, ale na pewno nie są to zastosowania jedyne, być może w kolejnym akapicie jest wymienione pięć

innych dziedzin, tak więc szczególna pozycja kryminalistyki jest tutaj całkowicie przypadkowa. Pytanie to można uratować modyfikując je do postaci "Do czego wykorzystuje się hipnozę w kryminalistyce?", czyli czyniąc je precyzyjniejszym, oraz odpowiednio modyfikując zakres odpowiedzi (“w procesie identyfikacji sprawców”).

2. Wielość "istot rzeczy"

> *Mleko stanowi podstawowy produkt do wyrobu różnych napojów mlecznych i serów.*
>
> Czym jest mleko?

Mleko jest bardzo wieloma rzeczami: jest produktem przemysłu mleczarskiego, jest wydzieliną gruczołów mlecznych, jest podstawą diety cieląt, jest cieczą. Część z tych określeń “istoty” mleka układa się w hierarchie pod kątem ogólności: mleko jest napojem<pokarmem<rzeczą jadalną. To, że otrzymaliśmy paragraf w którym mleko jest ujmowane jako produkt, a nie napój, lub wydzielina jest kwestią przypadku. Ponadto, często tego typu pytania są banalne.

> Toyota Corolla II stała się drugim najlepiej sprzedającym się samochodem 1970 roku na świecie.
>
> Czym jest Corolla?

Czasami jednak z różnych względów, wielość tych możliwych określeń jest zawężona, np. pytając “Kim jest X?” osoba X może być wieloma rzeczami naraz: synem, mężem, diabetykiem itd. Ale w wypadku pytań o ludzi, utarło się uznawać je za jednoznaczne z pytaniem o zawód, lub funkcję społeczną danej osoby. Takie pytania są dopuszczalne, bo nie są tak wieloznaczne jak pytania wyżej.

3. Wielość cech

> *Jest piątym co do wielkości naturalnym satelitą w Układzie Słonecznym.*
>
> Jaki jest księżyc?

Przecież księżyc ma ogromną liczbę różnorakich cech, jest biały, jest duży, składa się ze skał, jest widoczny nocą, to że autor wspomina akurat o tej cesze, nie świadczy o jej priorytetowym znaczeniu. W tym jaskrawym wypadku, mówimy nawet nie o cesze samego księżyca, tylko księżyca jako elementu zestawienia z innymi obiektami. To że tę cechę jesteśmy w stanie wskazać, jest wyłącznie efektem takiej a nie innej kompozycji tekstu, więc nie jest to dobre pytanie (i odpowiedź).

Co innego, kiedy pewna cecha jest akcentowana w sposób uzasadniony i świadomy:

Wieprzowina, dzięki łatwości rozrodu świń, stała się wkrótce najpopularniejszym mięsem.

Co wyróżnia świnie jako zwierzęta gospodarskie?

Pytania metatekstualne

Do kategorii pytań nadmiernie zależnych od tekstu możemy zakwalifikować także pytania metatekstualne, czyli odnoszące się nie do **tematu**, a bardziej do samego **tekstu jako tekstu**, albo języka w nim wykorzystanego.

...sytuację taką określa się jako „spodziectwo bez spodziectwa".

Czy forma "spodziectwa" jest prawidłowa?

Zamiast o schorzenie, pytamy o poprawność językową, wykorzystując tekst jako przykład normatywnego zastosowania języka. Jest to niepoprawne. Pytania powinny dotyczyć tematu, o którym jest tekst.

Oczywiście jeżeli tekst akurat dotyczy języka (np. dostaliśmy do anotacji artykuł o rzeczownikach), to pytania będą o języku, natomiast nie będą dotyczyć warstwy językowej samego artykułu.

Pytania o opinię

Czasami w tekście pojawiają się opinie jakichś ludzi, generalnie unikajmy układania pytań tak, żeby odpowiedź **polegała** na poprawności czyjejś opinii.

Platon nie chciał wtajemniczać zbyt młodych ludzi do spraw republiki, ponieważ uważał, że mają zbyt wiele zapału i są skłonni reformom.

Do czego dążą młodzi ludzie w polityce?

Nawet jeśli pogląd Platona jest prawdziwy, i młodzi ludzie faktycznie są mniej konserwatywni, tutaj taka teza nie jest stwierdzana, lecz tylko przytaczana. Lub bardziej ekstremalnie odnosząc się do kwestii które w ogóle nie dotyczą faktów, lecz ocen subiektywnych:

Krytycy uznali debiut reżysera za mało imponujący.

Czy pierwszy film był dobry?

Co innego gdy pytamy explicite o czyjąś opinię, i.e. nie uznajemy jej za miarodajną i reprezentatywną, lecz wyraźnie chcemy zapytać o to co ktoś powiedział lub sądził, wówczas takie pytanie jest na miejscu.

Piłsudski uważał że sprzymierzenie się z Ukraińcami pozwoli na zabezpieczenie wschodniej granicy.

Co zdaniem Marszałka miał przynieść sojusz z Ukrainą?

Umiejętność rozróżniania wypowiedzi o świecie, od cytatów na temat tego samego tematu jest cenna, więc tego typu pytania, proszące o przytoczenie czyjejś opinii są wartościowe

jako materiał szkoleniowy dla naszych modeli. Aby uzyskać tę umiejętność konieczne jest aby w zbiorze danych nie było pytań, traktujących opinie jedynie cytowane jako fakty. Pytania które odnoszą się do opinii tylko przytaczanych, są dobre jako pytania nieodpowiadalne.

Odpowiedzi

Do każdego pytania musi być możliwość podania na podstawie tekstu odpowiedzi. Wyróżniamy dwie warstwy anotacji odpowiedziami:

1. minimalny fragment występujący w tekście, wyrażający odpowiedź na pytanie
2. odpowiedź w “formie bazowej”

Relacje między warstwami

Warstwy te zostaną wykorzystane niezależnie od siebie, a więc każda z nich ma być pełnoprawną odpowiedzią na pytanie. Nie może być sytuacji, w której warstwa nr 2 jest komentarzem, który dopiero rozjaśnia nam, dlaczego zakreślony fragment miałby stanowić odpowiedź na pytanie, jak w poniższym przykładzie:

> *Pierwotnie Henryk miał się z nią rozwieść i oddalić ją, jednak na nieszczęście Katarzyny został przechwycony jej list miłosny do Culpepera.*
>
> O co oskarżono Katarzynę?
> Forma bazowa: *o zdradę*

Jeżeli nie jesteśmy w stanie znaleźć dla naszego pytania takiego samodzielnego fragmentu wyrażającego odpowiedź, nie jest to dobre pytanie do tego tekstu. W powyższym tekście nie ma mowy o oskarżeniu wobec Katarzyny, które doprowadziło ostatecznie do jej śmierci, tylko o podstawie tego oskarżenia. Lepsze byłoby więc takie pytanie:

> Co było dowodem zdrady, której dopuściła się Katarzyna?
> Forma bazowa: list miłosny do Culpepera

Oznacza to, że fragment powinien **wyrażać** odpowiedź na pytanie, a nie tylko ją sugerować. Nie powinno być też takiej sytuacji, w której zakreślany jest większy fragment (i.e. zawierający zbędne elementy które nie wyrażają odpowiedzi), w celu lepszego uzasadnienia, iż faktycznie fragment pasuje do pytania. Kierowanie się rozumowaniem, że "dopiero w tym szerszym kontekście, staje się jasne, że zakreślony fragment (często tylko jedno słowo lub dwa), faktycznie odpowiada na pytanie" jest błędne. Najczęściej, gdybyśmy kierowali się takim rozumowaniem, i tak musielibyśmy zakreślić cały paragraf, albo sporą jego część, żeby było wiadomo na pewno, na jaki jest temat.

> *Pod koniec VII wieku p.n.e. Nabuchodonozor II rozpoczął przesiedlenia ludności judejskiej na tereny Babilonu, zaś w roku 587 p.n.e. zburzył Jerozolimę, świątynię i uprowadził pozostałych Judejczyków.*
>
> Kiedy zniszczono Świątynię Jerozolimską po raz pierwszy?

Zaznaczony fragment jest tylko pewną datą, i wyrwany z kontekstu nie wskazuje na to, że faktycznie w tekście mowa jest o Świątyni Jerozolimskiej, natomiast nie jest to problemem. Interesuje nas tylko fragment wyrażający datę, bez konieczności zaznaczania późniejszej części zdania, która wspomina o świątyni.

Forma bazowa odpowiedzi

Warstwa nr 2 powinna powstać na podstawie anotacji z warstwy nr 1, i powinna stanowić taką wariację formy występującej w tekście, która brzmiałaby naturalnie jako odpowiedź udzielona w konwersacji na zadane pytanie. Najprostszym przypadkiem jest tu modyfikacja fleksyjna, na przykład:

> *O ile wczesna faza twórczości Kieślowskiego charakteryzowała się ascezą formalną na wzór dokumentów, o tyle późna cechowała się wyszukanymi zabiegami wizualnymi.*
>
> Na którym etapie działalności reżyserskiej Kieślowski charakteryzował się oszczędnością stylistyczną?
> Forma bazowa: wczesnym

W tym wypadku konwertujemy mianownikową do postaci miejscownikowej. Często będzie pewna dowolność pomiędzy wyborem formy mianownikowej i innego przypadka.

> *Hajle Syllasje I objął tron Etiopii w roku 1930.*
>
> W którym państwie rządził Hajle Syllasje?
>
> Forma bazowa: Etiopia (lub: w Etiopii)

W ramach przygotowywania form bazowych, możemy również rozwijać niektóre skróty (o ile rozwinięcie takie można wywnioskować z fragmentu, i nie są skrótami powszechnie rozpoznawalnymi, jak np. ONZ).

> *Rozkazem MSWojsk. nr 1921 z 23 maja 1919 przemianowano dywizję gen. Śmigłego-Rydza na 1 Dywizję Piechoty Legionów.* [...] *Twarda postawa 1 DP Leg., zaprezentowana podczas kampanii wrześniowej, wzbudziła szacunek Niemców, którzy określali ją mianem "żelaznej dywizji".*
>
> Która jednostka była nazywana “żelazną dywizją”?
> Forma bazowa: 1 Dywizja Piechoty Legionów

W formie bazowej możemy pomijać słowa nieistotne dla samego brzmienia odpowiedzi, które w tekście pojawiają się w charakterze wtrącenia:

> *Na początku 1915 istniały cztery okręgi: warszawski (Adam Koc, Bogusław Miedziński, ps. „Świtek”), lubelski (Andrzej Turczyński-Brenner, ps. „Mieczysław II” i Tadeusz Herfurt ps. „Armak”, radomski (Karol Rybasiewicz) i siedlecki (Bogusław Miedziński, Józef Korczak ps. „Piotr”).*
>
> Jakie okręgi tworzyły Polską Organizację wojskową w 1915?

Forma bazowa: warszawski, lubelski, radomski i siedlecki

W wypadku pytań o rozstrzygnięcie, forma bazowa musi przyjmować postać “Tak” lub “Nie”.

Minimalność fragmentów

Warstwa odpowiedzi w tekście, powinna składać się z minimalnych fragmentów wyrażających odpowiedź.

> *Zaręczyny zerwano tuż przed ślubem – 11 czerwca 2001. Powodem były odmienne wyznania narzeczonych* [...]
>
> Kiedy odwołano plany zawarcia ślubu?

Dla powyższego pytania zakreśliliśmy cztery fragmenty, które wydają się naturalnymi kandydatami na odpowiedź na pytanie. Fragmenty te zawierają się w sobie, i.e. fragment czerwony zawiera również to, co zakreślono pomarańczowym, żółtym i zielonym, fragment żółty zawiera też fragment zielony. Fragment czerwony, jest stanowczo zbyt wąski, w odpowiedzi interesuje nas czas zerwania zaręczyn, najlepiej żebyśmy nie musieli tej informacji dodatkowo wyłuskiwać z zaznaczonego fragmentu, tylko aby zrobił to za nas wytrenowany na przygotowanych danych model. W tym wypadku zdanie jest krótkie, ale oczywiście można sobie wyobrazić zdanie długie na kilka linijek, albo takie, w którym pojawia się kilka dat. Gdybyśmy chcieli zaznaczać za każdym razem całe zdanie, prowadziłoby to do niejednoznaczności. Dlatego generalnie (poza wyjątkami które opiszemy niżej), zaznaczanie całych zdań jest błędem.

Trzy pozostałe fragmenty, różnią się ilością przekazanej informacji. Fragment pomarańczowy jest nieco bardziej precyzyjną wersją fragmentu żółtego. Oba te fragmenty dodają zaś względem zielonego, informację, że zdarzenie to odbyło się niedługo przed zaplanowanym ślubem. W tym wypadku, takie okoliczności są raczej mało istotne, więc nie powinny znaleźć się w zaznaczonym fragmencie. Fragment zielony jest anotacją wzorcową.

> [...] *Ślub kościelny odbył się 2 maja w Senlis (departament Oise), rodowej siedzibie dynastii Kapetyngów, gdzie w 987 roku Hugo Kapet został wybrany pierwszym królem z tej dynastii.*
>
> Gdzie miała miejsce ceremonia zaślubin?

W tym wypadku stosujemy analogiczne znaczenie kolorów fragmentów. Informacja dodawana we fragmencie pomarańczowym (i uzupełniana we fragmencie czerwonym), miała na pewno istotne znaczenie symboliczne dla samego wydarzenia, natomiast nie jest w żaden sposób oczekiwana w pytaniu, więc jest nadmiarowa względem węższych fragmentów żółtego i zielonego. W tym wypadku oba te fragmenty mogą zostać uznane za poprawne (informacja o departamencie może być konieczna, jeżeli miejscowości o nazwie Senlis jest więcej niż jedna).

Decydując o tym, które elementy są istotne, a które nadmiarowe,

> *Stoczył tu dwie zwycięskie bitwy – pod Smoleńskiem i Borodino, jednak rosyjski feldmarszałek Michaił Kutuzow kontynuował odwrót, stosując jednocześnie taktykę spalonej ziemi, a Napoleon gnał Wielką Armię dalej na wschód*

Kto dowodził wycofującą się armią rosyjską?

W tym wypadku rozważamy cztery fragmenty. Fragment czerwony jest zbyt szeroki: wspominanie o “rosyjskości” Kutuzowa jest wręcz redundantne. Fragmenty pomarańczowy i żółty można uznać za równo wartościowe. Stopnie wojskowe, pozycje społeczno-polityczne (*prezydent*), stałe profesje (*malarz Paul Gauguin*) i podobne, można uznać za stałe i istotowe atrybuty ludzi, więc zaznaczanie ich nie jest redundantne. Idąc w kierunku minimalizmu nie należy pomijać informacji istotnych, np. mimo iż często odnosimy się do znanych osób tylko po nazwisku, w tym wypadku występuje również imię, więc nie ma powodów by eliminować tę istotną część ”nazwy” danego “przedmiotu”, i stosować anotacji oznaczonej kolorem purpurowym.

Zaznaczanie zdań

Istnieją sytuacje, w których poprawną procedurą jest oznaczenie całego zdania. Są to przede wszystkim:

1. Pytania o rozstrzygnięcie.
2. Pytania, odpowiedzią na które jest opis jakiejś sytuacji, zdarzenia, faktu.

W pierwszym wypadku, poprawną anotacją jest zaznaczenie zdania, albowiem dopiero ono posiada *wartość logiczną*, a więc dopiero zdanie (a nie tylko nazwa, lub pojedyncze określenie) może wskazywać odpowiedź “Tak”, lub “Nie”.

> *W końcu Hitler namówił Heßa do zawarcia w 1927 roku związku małżeńskiego, aby położyć kres pogłoskom o jego homoseksualizmie.*
>
> Czy Hess był żonaty?
> Forma bazowa: Tak

Zaznaczamy wyrażenie posiadające podmiot oraz orzeczenie, nie jest to jednak pełne zdanie, lecz jedno ze zdań zdania złożonego, pomijające okolicznik “W końcu”. Spełnia jednak kryteria zdania, mogłoby funkcjonować w innym tekście jako pełne zdanie, oraz posiada wartość logiczną, więc może być podstawą do udzielenia odpowiedzi “Tak”.

W drugim wypadku wymóg taki wynika z natury spodziewanej odpowiedzi, np.

> *Wylądował w pobliżu Floors Farm, szesnaście kilometrów od centrum Glasgow w Szkocji. Został znaleziony przez rolnika Davida McLeana, który zaprowadził go do swojego gospodarstwa. Zabrali go stamtąd funkcjonariusze Gwardii Krajowej. Podczas przesłuchania, przedstawił się jako Alfred Horn.*
>
> Jak Rudolf Hess dostał się do niewoli?

W tym wypadku musimy podać sposób dostania się w ręce Brytyjczyków, a więc pewien proces, *ergo* przytoczyć zdanie, a nawet (jak w tym wypadku) dwa zdania.

W powyższych sytuacjach, należy pamiętać o tym, aby zdanie zawierało 1. podmiot, 2. orzeczenie. Taki wymóg może być rozluźniony, jeżeli zdania są bardzo długie, i będzie to prowadzić do nadmiernego rozrostu zakreśleń, np. tutaj pomijamy podmiot (“Napoleon I”)

Napoleon I uważając, że wygrać z Wielką Brytanią może tylko na drodze jej blokady, powziął wówczas zamiar pobicia Rosji i zmuszenia jej do respektowania zobowiązań w Tylży - uderzył nie znając jednak znaczenia przestrzeni i klimatu Rosji.

Co było powodem klęski Napoleona w Rosji?

Poprawnymi anotacjami będą także równoważniki zdań.

Wielokrotne występowanie odpowiedzi w tekście

Jeżeli odpowiedź pojawia się w tekście wielokrotnie, starajmy się wybrać to wystąpienie, którego bezpośredni kontekst najściślej łączy się z pytaniem.

Podstawowym źródłem pisanym dla wczesnej historii Żydów jest Biblia, która określa ich jako naród wybrany przez Boga Jahwe. Żydzi mieli pochodzić od Abrahama, który około 1800 r. p.n.e. przywędrował z Ur w Mezopotamii do ziemi Kanaan (obecnie Palestyna oraz państwo Izrael). Podczas trójpokoleniowego pobytu w Kanaanie Żydzi zaczęli posługiwać się językiem starokananejskim (będącym punktem wyjścia do języka hebrajskiego),

Który naród ma według Biblii pochodzić od Abrahama?