Magyar nyelvű időjárásjelentések nyelvi modell alapú automatizált generálása

Knap Árpád[1], Dömsödi L Bíborka[2], Mogyorósi Pálma[2],
Szigeti Péter[3], Tóth Andor[3], Virág Attila[3] és Kmetty Zoltán[1]

[1] Társadalomtudományi Kutatóközpont
{knap.arpad, kmetty.zoltan}@tk.hu
[2] iStat Consulting Kft.
[3] Central Médiacsoport Zrt.

Kivonat: Tanulmányunkban bemutatjuk egy olyan eszköz első változatát, amely képes az Országos Meteorológiai Szolgálat adatai alapján magyar nyelvű időjárásjelentések automatizált megírására. Az eszköz első lépésben egy szabályalapú rendszerben a nyers numerikus adatokból előállít egy listát a várható időjárási eseményekről, majd ezen listából az OpenAI GPT-3 nyelvi modelljére támaszkodva a megfelelő paraméterezés mellett kigenerál egy angol nyelvű időjárásjelentést, amit a DeepL Translate szolgáltatással magyar nyelvűre fordít. A kapott időjárásjelentések kifejezetten magas, 4 feletti értékelést kaptak a tesztelés során (1-5 skála) nyelvhelyesség, stílus és koherencia dimenziókban.

1 Bevezetés, célkitűzés

A természetes nyelv generálás (Natural Language Generation - NLG) nem számít új iránynak a számítógépes nyelvészeten belül, de az elmúlt években az egyre komplexebb tudású nyelvi modellek elterjedésével új lendületet kapott a terület. Az NLG-n belül két nagy kutatási és alkalmazási irány létezik: a szövegből szöveg generálás (text-to-text generation) és az adatból szöveg generálás (data-to-text generation) (Gatt és mtsai. 2018). Utóbbi irány elsősorban a robot újságírás kapcsán jelenik meg, különféle témaköröket érintve. Az egyik kiemelt terület a robot újságíráson belül az időjárásjelentések automatikus megírása (Goldberg és mtsai, 1994; Reiter és mtsai, 2005; Belz 2008; Turner és mtsai, 2007; Ramos-Soto és mtsai, 2014). Az időjárási hírek általában rövidek és jól felépíthetőek numerikus adatokból, ráadásul az adatok legtöbb esetben egy jól definiált és állandó struktúrában állnak rendelkezésre, ami jó alapot ad egy NLG projekthez.

A 24.hu, a TK[1] és az iStat Consulting egy közös projektet indított 2022 tavaszán, aminek az volt a célja, hogy magyar nyelvre előálljon az első olyan eszköz, amely automatikusan képes időjárásjelentéseket generálni OMSZ adatok alapján, nyelvi modell

[1] A publikációban szereplő kutatást, amelynek megvalósításában a Társadalomtudományi Kutatóközpont is részt vett, az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

segítségével. Összességében tehát magyar nyelvű, szöveges időjárásjelentéseket kívántunk automatizáltan előállítani, amelyhez mintaként a 24.hu meglévő időjárásjelentéseit tekintettük, célunk ezek reprodukálása volt.

A tanulmány első felében bemutatjuk a projektben használt kiinduló adathalmazokat, illetve az ezeken végrehajtott transzformációs lépéseket és logikai műveleteket, amelyek segítségével az adatokból kiválasztjuk a jelentések megírásához szükséges adatpontokat. Ezt követően ismertetjük az általunk használt nyelvi modell sajátosságait, illetve kiértékeljük a rendszer működését emberek által írt időjárásjelentések szövegeinek és általunk generált jelentéseknek az összehasonlításával. Végezetül ismertetjük a projekttel kapcsolatban felmerült kihívásokat és limitációkat, illetve továbbfejlesztési irányokat is megfogalmazunk.

Munkánk jelenlegi, első fázisában az adott napra vonatkozó (reggeli) időjárásjelentések generálását tűztük ki célul. Ezek a jelentések jellemzően az adott nap reggelétől a következő nap reggeléig tartó időszakra tartalmaznak adatokat. A jelentések előállításához kizárólag a 24.hu számára rendelkezésére álló OMSZ adatszolgáltatás keretében érkező adathalmazokat használjuk fel, és ezek közül is kizárólag azokat, amelyek számszerű adatokat, és nem szöveges jelentéseket tartalmaznak. Célunk volt, hogy a létrejövő jelentésekben megkülönböztessük egymástól az eltérő napszakokat, tehát ne az egész napra adjunk egy általános képet, hanem eltérő időjárási események esetében az egyes napszakokra különböző megállapításokat tegyünk. Szintén fontosnak tartottuk, hogy ezt az elvet ne csak a napszakok esetében, hanem földrajzilag is kövessük, ezért a jelentésekben, ha szükséges, megkülönböztetjük egymástól az ország különböző régióit. A jelentések előállításához az OpenAI GPT-3 (Radford és mtsai, 2018; Floridi és Chiriatti, 2020) angol nyelvi modelljét használjuk, amely szabadszöveges instrukciókat fogad, így az adatok feldolgozása, és szöveges paranccsá alakítása angol nyelven történik, a létrejövő jelentéseket pedig automatizáltan, a DeepL Translate szolgáltatással fordítjuk magyar nyelvre.

2 Adatfeldolgozás

Ahogy említettük, az OpenAI GPT-3 modellje szabadszöveges instrukciókat fogad bemenetként. Az alábbiakban bemutatjuk azt a folyamatot, amely során az OMSZ által közölt XML formátumú állományokból számos lépésen keresztül angol nyelvű szöveges instrukciókat hozunk létre.

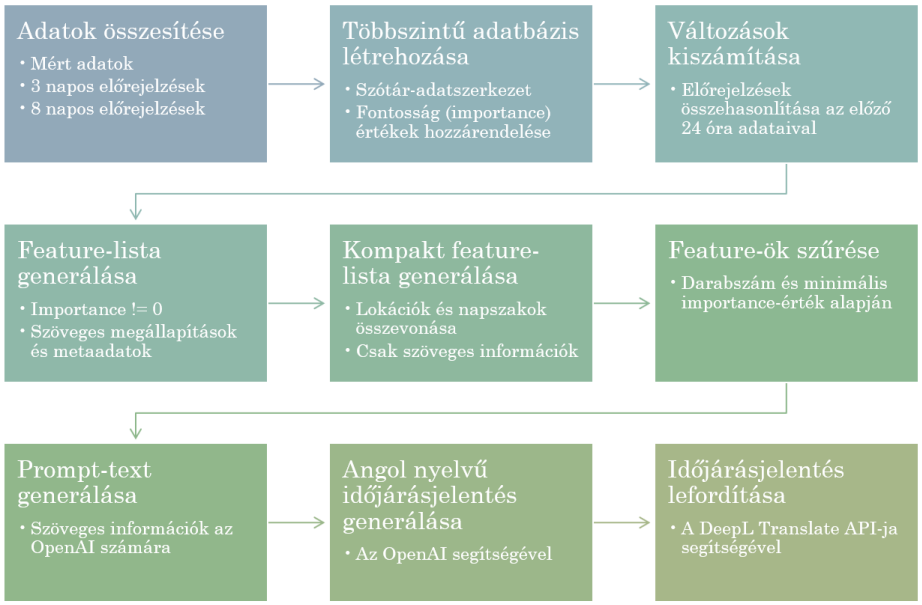

Fig. 1. az adatok transzformációjának és az időjárásjelentés megírásának folyamatábrája

2.1 Forrásadatok bemutatása

A 24.hu számára rendelkezésre álló OMSZ adatok közül három olyan állománytípust azonosítottunk, amely számszerű, tehát nem szöveges adatokat tartalmaz. Ezek az adatok (az adatformátum megváltozása miatt) 2022 január 10-e óta állnak rendelkezésünkre.

Kiinduló adataink első típusa a múltra vonatkozó tényadatokat tartalmaz. Ezekben az óránként kiadott XML állományokban egy-egy jelentés áll rendelkezésre 7 különböző településre (Budapest, Debrecen, Győr, Miskolc, Pogány, Sármellék, Szeged). Ami a fájlok adattartalmát illeti, időjárás kód (pl. derült, ködös, szitálás, eső, havazás), hőmérsékletérték, szélirány, szélerősség, légnyomás, továbbá relatív páratartalom érhető el az egyes településekre.

Kiinduló adataink második típusát a három napos előrejelzéseket tartalmazó fájlok jelentik. Ezeket minden nap két alkalommal bocsátja rendelkezésre az OMSZ: délben, valamint éjfélkor. Ezek az állományok óránkénti felbontású előrejelzéseket tartalmaznak, a következő három napra, 110 lokációra megadva. Az egyes adatpontok esetében hőmérsékletérték, illetve időjárás kód érhető el.

A harmadik kategóriába a nyolc napos előrejelzések tartoznak, amelyeket hasonlóan a három napos állományokhoz, szintén napi két alkalommal, délben, illetve éjfélkor adják ki, és ugyanarra a 110 helyszínre tartalmaznak adatokat. Ezekben a fájlokban a következő nyolc napra láthatunk előrejelzéseket, amelyek nem óránkénti, hanem napi felbontásúak, tehát a következő nyolc napra egy-egy adatpontot találunk. Ezeknek a

fájloknak az adattartalma is eltérő, itt 24 órás maximum és minimum hőmérséklettel, csapadékmennyiséggel, maximum szélerősséggel, továbbá csapadékvalószínűséggel dolgozhatunk.

2.2 Adatranszformációs lépések és logikai műveletek

Ahhoz, hogy az XML formátumú adathalmazokból eljussunk az OpenAI modellje számára inputként beadandó szabadszöveges instrukciók előállításáig, számos adattranszformációs lépést hajtunk végre. Elsőként az adatainkat táblázatos formátumúra hozzuk, majd ezt követően egy szótárszerkezetben tároljuk el őket. Ezután a szótárszerkezetből egy lapos listát hozunk létre, amelyben már megjelenik az adott adatpont szövegszerű reprezentációja. Ennek a listának a további transzformációjával állnak elő a szöveges adataink, amelynek szűrésével létrejön az az angol nyelvű prompt-szöveg, amelyet inputként használunk az OpenAI nyelvi modellben.

2.2.1 Az adatok összesítése

A folyamat a mért adatok összesítésével indul. Ennek során az a célunk, hogy az időjárásjelentés generálásához képest vett elmúlt 24 óra adatait összegezzük. Ehhez az adatokat az XML állományokból táblázatos formátumúra konvertáljuk, és napszakokra[2], valamint napszak/nagyrégió[3] bontásban aggregáljuk, kiszámítva a hőmérséklet és a szélerősség maximális, minimális, átlagos és medián értékeit, illetve az adott napszakra vonatkozó időjárás kódok, valamint az esetleges adathiányok darabszámát (pl. 6db borult, 1db szitálás, 21db eső, 2db zápor, 4db adathiány).

A következő lépésben a három napos előrejelzéseket összegezzük. Ennek során az adott napra, illetve a következő éjszakára vonatkozó adatokat dolgozzuk fel. Az adatokat napszakonként aggregáljuk négy különböző földrajzi felosztásban – nagyrégiókra, régiókra, megyékre, illetve a 110 lokációra, kiszámítva az adott napszakra vonatkozó maximális, minimális, átlagos, medián hőmérsékletet, valamint az időjárás kódok számát.

Ezt követően a nyolc napos előrejelzéseket dolgozzuk fel, hasonlóan a három napos állományokhoz. Mivel itt a kiinduló fájlok napi felbontásban tartalmaznak információkat, nem különböztetjük meg egymástól az egyes napszakokat, tehát az adott napra aggregáljuk adatainkat, nagyrégió, régió, megye, illetve lokáció bontásban. A folyamat során kiszámítjuk a hőmérséklet, csapadékmennyiség, szélerősség és csapadékvalószínűség adatok különböző szélső- illetve középértékmutatóit.

Az adatok összesítésével párhuzamosan elvégzünk néhány további műveletet is. Egyebeket mellett például egyszerűsítéseket végzünk az időjárás kódokon, pl. az ónos

[2] A projektben ötféle napszakot különböztetünk meg egymástól: 0:00-tól 6:00-ig éjszaka, 6:00-tól 9:00-ig reggel, 9:00-tól 12:00-ig délelőtt, 12:00-tól 18:00-ig délután, 18:00-tól 0:00-ig pedig este.

[3] Magyarországon a következő nagyrégiókat (NUTS-1 szint) különböztetjük meg egymástól: Dunántúl, Alföld és Észak, Közép-Magyarország.

szitálást és az ónos esőt egyaránt ónos esőre kódoljuk, az egyes mérési pontokhoz pedig hozzárendeljük a megyét, a régió, illetve nagyrégió elnevezéseket. Ezen kívül a szélerősségértékekhez hozzárendeljük a Beaufort-skála egy egyszerűsített verzióját[4].

2.2.2 Szótár létrehozása és fontosság értékek hozzárendelése

Ezzel előálltak a táblázatos adataink – a következő lépés annak eldöntése, hogy az egyes adatpontok közül melyek azok, amiket fel kívánunk használni az aktuális napi időjárásjelentés megírásához. Ennek eldöntéséhez egy többszintű szótárat (Python dictionary-t) használunk, amiben minden adatponthoz hozzárendelünk egy-egy fontosság (importance) értéket. A fontosság érték 0 és 1 között mozog, ahol az 1 nagyon fontos (kötelezően szerepelnie kell a jelentésben), a 0 pedig az egyáltalán nem fontos. A fontossági értékek bevezetésének az a célja, hogy az előálló időjárásjelentés ne legyen indokolatlanul hosszú, és csak olyan információkat tartalmazzon, amelyek érdemben elmondanak valamit az adott napról.

Az előálló szótár-adathalmaz kétféle típusú adatot tartalmaz: (1) egyrészt pusztán az előrejelzésekből származó adatokat, másrészt (2) a mért adatok és az előrejelzések összehasonlításán alapuló változásokat. Ami az első típust illeti, ide tartoznak a legalacsonyabb éjszakai és esti hőmérsékletre, valamint a legmagasabb nappali hőmérsékletre vonatkozó adatok, a különböző napszakokra vonatkozó időjáráskódok eltérő földrajzi bontásokban, továbbá a csapadékra és szélre vonatkozó adatok szintén különböző földrajzi bontásokban. Az előző napi mért adatok és az előrejelzések összehasonlításán alapuló adatok között jelenleg 12 féle megállapítást különböztetünk meg egymástól. Ilyen például az, ha derült, napos idő után felhős, borult időre számíthatunk, ha jelentős felmelegedés vagy lehűlés várható, vagy ha tartósan derült, esetleg borult idő valószínű.

2.2.3 Ismétlés helyett összevonás

A szótár adathalmaz létrehozásához és a fontosság értékek hozzárendeléséhez különböző logikai döntések meghozatalára van szükség. Az első lépés annak a megállapítása, hogy mik azok az adatok, amik „kötelezőek”, tehát mindenképp kellenek egy időjárásjelentésbe. Ezek közé tartozik például a legalacsonyabb éjszakai és a legmagasabb nappali hőmérséklet. Ezek 1-es fontosság értéket kapnak és a teljes országra vonatkoznak.

A következő lépés az egyes időjárási események fontossági kategóriákba való besorolása. Ezeket az időjárás kódok jelölik. Itt háromféle kategóriát különböztetünk meg egymástól. (1) A rendkívüli események 1-es fontosság értéket kapnak amennyiben megjelennek, függetlenül attól, hogy milyen arányban jelennek meg az adott bontásban (a következő időjárás kódok tartoznak ide: ködös, zápor, zivatar, ónos eső, hószállingózás, havazás, hózápor, havaseső, hózivatar, hófúvás). (2) A nem csapadékra, hanem derült vagy felhős időre vonatkozó kódok közül az adott bontásban a domináns kódot keressük, tehát azt, ami legalább egyharmados arányban jelenik meg (a következő kódok tartoznak ide: derült, kissé felhős, közepesen felhős, erősen felhős, borult). Ha több ilyen kód is van, akkor felsoroljuk ezeket a kódokat, és azt a megállapítást tesszük

[4] Ez a skála az eredeti 13 helyett 6 fokú.

például, hogy „kissé, illetve közepesen felhős idő várható". Amennyiben nincsen domináns kód, de több is megjelenik a felsoroltak közül, akkor azt mondjuk, hogy változóan felhős idő várható[5]. (3) Ezeken kívül a szitálást és az esőt különböztetjük meg. Az erre vonatkozó kódoknak legalább 20 százalékos arányt kell elérniük az adott bontásban, hogy nullától eltérő fontossági értéket kapjanak.

A fontosság értékek meghatározásakor általánosságban azt az elvet követjük, hogy a legnagyobb fontosság értéket a rendkívüli időjárási eseményre vonatkozó kódok kapják, ezt követik a csapadékra vonatkozó egyéb kódok, utána pedig a nem csapadékra vonatkozó kódok következnek. Tehát abban az esetben, ha például a Dunántúlon kissé felhős idő lesz, akkor az erre vonatkozó megállapítás 0,4-es fontosság értéket kap, ha viszont eső várható, akkor az 0,7-es értéket. Ezen kívül a teljes országra vonatkozó megállapítások nagyobb fontosságot kapnak, mint a nagyrégiós és régiós, a nagyrégióra vonatkozó megállapítások pedig nagyobb fontosságot kapnak, mint a régiós bontású előrejelzések.

A logikai döntések következő kategóriája az országos, nagyrégiós és régiós felbontású adatok kezelésére vonatkozik. A cél az, hogy eltérő régiókra vonatkozó, de azonos tartalmú előrejelzések ne jelenjenek meg ismételten a jelentésekben, mivel ezek nagyon hosszúvá és repetitívvé tennék a keletkező szövegeinket. Itt azt az elvet követtük, hogy először az országos szintet vizsgáljuk az egyes napszakokra. Ha egy adott időjárási kód az előzőekben ismertetett logika szerint itt megjelenik, akkor 1-es fontosság értéket kap az adott napszakra. Ezután azt vizsgáljuk, hogy ez az esemény az adott napszakban hogyan jelenik meg a nagyrégiókban. (1) Ha mindhárom nagyrégióban megjelenik, akkor az egy országos esemény. Ilyenkor nincsen további teendőnk, mert országosnak jelöltük az eseményt. (2) Ha két nagyrégióban jelenik meg, akkor az nem egy országos esemény. Ilyenkor a releváns nagyrégiókhoz hozzárendeljük a nullától eltérő fontosság értékeket, és az országos szinten pedig nullát állítunk be. (3) Ha egy nagyrégióban jelenik meg, akkor az szintén nem egy országos esemény. Ilyenkor megvizsgáljuk azt, hogy az adott nagyrégió alá tartozó kisebb régiók közül hányban jelenik meg az adott esemény. Ha nem az összesben, akkor ezekhez a kisebb régiókhoz hozzárendeljük az fontosság értékeket, és az országos szinthez pedig a nulla értéket.

A csapadékkal kapcsolatos adatokat hasonlóan kezeljük, azzal a különbséggel, hogy mivel ezek a nyolc napos, napi felbontású előrejelzésekből származnak, nem bontjuk őket napszakokra. Mivel viszont az az információ rendelkezésünkre áll az időjárás kódokból, ha valahol csapadék várható, a csapadék adatokat arra használjuk fel, hogy a jelentős mennyiségű várható csapadékot előrejelezzük. A jelentős mennyiség alsó határát napi 10mm-ben definiáltuk.[6]

A következő lépésben a széllel kapcsolatos adatainkat összegezzük. A szélre vonatkozó előrejelzésekhez a fontosság értékeket az általunk létrehozott, hatfokozatú, egyszerűsített Beaufort-skála alapján rendeljük hozzá az adatpontokhoz. Ennek segítségével a széllökések maximális erősségét soroljuk be. Természetesen a magasabb

[5] Ezt csak abban az esetben tesszük, ha összességében ezek a kódok tényleg dominálnak, tehát legalább 50%-os arányt érnek el.

[6] Ahogyan az eddigiekből is látszik, a program működésének logikája tovább finomítható, bonyolítható. A cikkünkben tárgyalt első verzióban az említett határértékek rugalmasan változtathatók, emellett alkalmazásuk jelentősen egyszerűsítette például a jelentős mennyiségű csapadék előrejelzésének problémáját.

Beaufort-érték magasabb fontosság értéket eredményez. (1) Országos szintű adatként a (kis)régiókhoz hozzárendelt értékek móduszát, tehát a leggyakoribb értéket fogadjuk el. Ilyen módon elkerüljük az országos szélintenzitás felülbecslését. (2) Ha két nagyrégióban az országosnál magasabb Beaufort értéket látunk, akkor ezekhez a régiókhoz hozzárendeljük az adott fontosság értéket. (3) Amennyiben egy nagyrégióban, és azon belül több mint két régióban az országosnál magasabb Beaufort értéket látunk, akkor az adott nagyrégióhoz rendelünk hozzá nullától eltérő fontosság értéket. (4) Végezetül, ha egy nagyrégióban, és azon belül legfeljebb két régióban látunk az országosnál magasabb Beaufort értéket, akkor az adott régiókhoz rendeljük hozzá a megfelelő fontosság értékeket.

A logikai műveletek utolsó köre a napszakok kezelésére vonatkozik. Ennek fő célja – a régiókhoz hasonlóan – annak elkerülése, hogy ugyanazokat az eseményeket számos régió/napszak kombinációra újra és újra leírjuk a jelentésben, mivel az nagyon ismétlődővé és hosszúvá tenné a szöveget. Általánosságban azt az elvet követjük, hogy legfeljebb két napszak esetén megnevezzük a napszakokat. Ha az esemény ennél több napszakra igaz, akkor az egész napra teszünk megállapítást.

2.2.4 Úton a szövegszerű adatok felé

A következő lépésben a többlépcsős fa adatszerkezetből egy „lapos" attribútum-listát készítünk. A fenti logikai műveletek egy részét is ezen konverzió során hajtjuk végre. Ez a formátum egy köztes szintet képez a többszintű adathalmaz és a nyelvi modell számára előállítandó szöveges instrukciók között. Ezt azt jelenti, hogy itt még vannak számszerű adataink, de már megjelennek a szöveges adatok is. Az attribútum-lista már egy szűrt adatkészletet tartalmaz: ebbe csak a nullától eltérő fontosság értékkel rendelkező adatpontok kerülnek bele.

Ezt követően tovább egyszerűsítjük az adatszerkezetet, amelynek az eredményét kompakt attribútumoknak neveztünk el. Itt már kizárólag a szöveges információk és a hozzájuk tartozó fontosság értékek jelennek meg. Ebben a fázisban történik meg a napszakok és a régiók összevonása „egy sorra". Ennek az a célja, hogy például a különböző napszakokra vonatkozó, de megegyező tartalmú megállapítások egy instrukcióként kerüljenek be a modellbe.

Ezt követően a keletkezett listát tovább szűrjük az alapján, hogy hány darab adatpontot szeretnénk felhasználni az adott időjárásjelentés megírásához. Ez a folyamat egy függvény segítségével rugalmasan változtatható kimenetellel hajtható végre aszerint, hogy milyen hosszú, mennyire részletes időjárásjelentést szeretnénk generálni. A szűrést az alábbi elvek alapján hajtjuk végre. (1) Az 1-es fontosság értékkel rendelkező adatok mindenképp megjelennek, akkor is, ha kevesebb attribútumot kértünk, mint ahány ilyen adatpont van. (2) A fennmaradó darabszámú helyre fontosság szerint csökkenő sorrendben kerülnek be a megállapítások. (3) A megegyező fontosságú megállapítások véletlen sorrendben kerülnek be. (4) A lista összeállításakor meghatározhatunk minimális fontosság értéket is.

Utolsó lépésként előállítjuk azt a szöveges instrukciót, amelyet az OpenAI GPT3 modellje számára beadunk, mint inputot – ezt a szöveget prompt-szövegnek neveztük el. Ez tartalmazza azokat a megállapításokat, amelyek a fenti lépések eredményeképpen előálltak. A prompt-szöveg minden esetben az alábbi mondattal kezdődik:

```
write a weather report for today using the following
information
```[7]

Ezt követően a leszűrt attribútumok felsorolása olvasható, vesszővel elválasztva. Fontos, hogy a prompt szövegben megadott adatok sorrendje visszaköszön a generált jelentésben is, emiatt érdemes ezeket olyan sorrendben megadni, ahogyan azok általában a valós jelentésekben is szerepelnek – mi is ezt a logikát követtük a projektben. Az 1. táblában egy példa olvasható egy prompt szöveggel és egy ebből előállt időjárásjelentéssel.

1. táblázat: prompt szöveg és az OpenAI GPT-3 modellje által generált időjárásjelentés magyarra fordított változata

Prompt	Generált időjárásjelentés	Magyarra fordított jelentés
lowest night temperature between -1.0 and 7.0 degrees, daytime temperature between 8.7 and 20.3 degrees, rain possible, very cloudy weather in the evening and night, thunderstorms possible in the Southern Great Plain and Northern Great Plain regions, sleet possible in the night in the Western Transdanubia and Southern Transdanubia regions, snow possible in the night in the Western Transdanubia region, large amounts of precipitation is possible in the Great Plain and North and Transdanubia regions, the weather warms up significantly, strong gale expected in the Great Plain and North and Transdanubia regions	The lowest night temperature will be between -1.0 and 7.0 degrees, and the daytime temperature will be between 8.7 and 20.3 degrees. It is possible that it will rain, and the weather will be very cloudy in the evening and night. There is a possibility of thunderstorms in the Southern Great Plain and Northern Great Plain regions, and sleet is possible in the night in the Western Transdanubia and Southern Transdanubia regions. Snow is possible in the night in the Western Transdanubia region. Large amounts of precipitation are possible in the Great Plain and North and Transdanubia regions, and the weather will warm up significantly. A strong gale is expected in the Great Plain and North and Transdanubia regions.	A legalacsonyabb éjszakai hőmérséklet -1,0 és 7,0 fok között, a nappali hőmérséklet pedig 8,7 és 20,3 fok között alakul. Elképzelhető, hogy esni fog az eső, este és éjszaka pedig erősen felhős lesz az idő. A Dél-Alföld és az Észak-Alföld térségében zivatarok kialakulására van kilátás, a Nyugat-Dunántúlon és a Dél-Dunántúlon pedig éjszaka havas eső, zápor is lehet. A Nyugat-dunántúli régióban éjszaka havazás lehetséges. Az Alföldön, valamint az Észak-Alföld és az Észak-Dunántúl térségében nagy mennyiségű csapadék hullhat, és jelentősen felmelegszik az idő. Az Alföldön és a Dunántúlon erős szélvihar várható.

[7] A kezdőprompt szövegének kiválasztásánál elsősorban az OpenAI dokumentációjában szereplő ajánlásokat tekintettük alapnak. A fentihez képest próbálkoztunk más, kissé eltérő megfogalmazású promptok alkalmazásával is, de a generált időjárásjelentések értékelése során ezt a megfogalmazást értékeltük a legjobbnak.

Ahogyan a fenti példában is látható, az elkészült magyar nyelvű jelentés jól olvasható, megfelelő magyarsággal íródott, tartalmi és szakmai szempontból elfogadható. Néhány apróbb hibára azonban felhívnánk a figyelmet. Ilyen például a régiók nevének kezelése, amely az Alföld és Észak (Great Plain and North) nagyrégió esetében nem megfelelően kerül lefordításra (a „Large amounts of precipitation are possible in the Great Plain and North and Transdanubia regions" mondatrészből "Az Alföldön, valamint az Észak-Alföld és az Észak-Dunántúl térségében nagy mennyiségű csapadék hullhat" lesz). Mivel az Alföld és Észak nagyrégió a köznyelvben kevésbé használatos, ezért érdemes lehet ehelyett a „keleti országrészben" szókapcsolat alkalmazása, amely egyébként is gyakrabban szerepel valós időjárásjelentésekben. Szintén megfontolandó a teljes országra vonatkozó adatok esetében ennek explicit módon való kiemelése. A kialakított rendszer kellően rugalmas ahhoz, hogy az ehhez hasonló finomításokat egyszerűen lehessen implemetálni a későbbiek során.

3 Időjárásjelentések generálása

Az időjárásjelentések generálása tehát az OpenAI API-ja segítségével történik. Jelenleg négy GPT-3-ra épített nyelvi modell érhető el az OpenAI felületén, az Ada, a Babbage, a Curie és a Davinci. Ezek a modellek árban, futási időben és teljesítményben is eltérnek egymástól. A modellek közül az Ada a leggyorsabb és legolcsóbb, de képességeit tekintve ez a legbehatároltabb, míg lista másik végén a Davinci van, ami lassabb és drágább, de képes bonyolult nyelvi feladatok megoldására is.

A jelentés megírásához jelenleg a text-davinci-002 modellt használjuk, mivel az alacsony token szám miatt ez a modell is gyorsan fut és alacsony költségen tartható (egy napi jelentés generálása kevesebb mint 10 centbe kerül). A modell széleskörűen paraméterezhető. Három olyan paramétert azonosítottunk, ami kifejezetten fontosnak bizonyult a szöveggenerálásnál. A „temperature" azt szabályozza, hogy mennyire legyen kreatív a nyelvi modell. Magasabb temperature értéknél olyan elemekkel gazdagítja a modell a szöveget, amik nincsenek benne a prompt szövegben, de kontextusában a szövegbe illenek. Például eső esetén belekerülhet az időjárásjelentésbe, hogy érdemes aznap esernyőt magunkkal vinni, ha kimegyünk az utcára. A „frequency penalty" paraméterrel büntethetjük a szóismétlést, míg a „presence penalty"-val azt szabályozhatjuk, hogy mekkora valószínűséggel jelenjenek meg új tokenek a szövegben.

Az időjárásjelentéseknél fontos az „adathűség", tehát nem szerencsés magasra állítani a temperature értéket, mert ezzel növeljük annak a kockázatát, hogy olyan részek is bekerülnek a szövegbe, amit az adatok nem támasztanak alá. Az általunk optimálisnak talált temperature paraméter esetében a tesztelés során azt tapasztaltuk, hogy a nyelvi modell nem ad hozzá olyan megállapításokat a szöveghez, amelyet a prompt szöveg nem tartalmaz. Ugyanakkor az is cél továbbá, egyben üzleti szempontból is indokolható, hogy ne legyen nagyon unalmas és repetitív az időjárásjelentés, tehát érdemes csökkenteni a szóismétléseket a szövegekben, hogy a látogatók szívesen olvassák azokat. A vonatkozó vizsgálataink alapján tehát úgy találtuk, hogy az alábbi paraméterek segítségével állíthatók elő a legjobb minőségű időjárásjelentések:

- temperature [0:1] = 0,45

- frequency_penalty [-2:2] = 0,5
- presence_penalty [-2:2] = 0

Az angol nyelvű jelentéseket a DeepL Translate API-ja segítségével fordítjuk végül magyarra. Alternatív megoldásként megvizsgáltuk a projekt keretében a Google Translate API-t használatát is. Utóbbi magyarról angolra jobban fordította az időjárás jelentéseket, de angolról magyarra a DeepL bizonyult megfelelőbbnek[8] a tesztjeink alapján.

3.1 A rendszer teljesítményének kiértékelése

A fenti folyamat eredményeképpen előálló időjárásjelentések minőségének kiértékelését 100 darab valós, a 24.hu által a rendelkezésünkre bocsátott, illetve 100 darab általunk generált időjárásjelentésen végeztük.[9] A jelentések generálásához a fent leírt OpenAI paramétereket használtuk, az attribútumok szűrése során pedig 10 darab adatpontot kértünk le, 0,5-ös minimális fontosság értékkel. A jelentéseket 100 darab véletlenszerűen kiválasztott napra generáltuk 2022 január 11. és 2022 július 20. között.

[8] A Google Translate tapasztalataink szerint nagyobb arányban fordította hibásan a különböző régióneveket tartalmazó mondatokat. Az 1. táblázatban közölt példa utolsó mondatát („A strong gale is expected in the Great Plain and North and Transdanubia regions”) hibásan a következőre fordította: „Erős szélvihar várható az Alföldön, valamint az Észak- és a Dunántúlon”.

[9] Ahogyan korábban is említettük, az általunk elérhető adatok köre és időintervalluma limitált volt, ezért nem állt módunkban ugyanazon napok esetében a valós időjárásjelentésekhez hasonlítani az általunk generált szövegeket. Az adatokon kívül további kihívást jelentett az, hogy nincsen információnk arról, hogy az újságíró vagy meteorológus pontosan melyik időpontban közzétett OMSZ modell, illetve az OMSZ által biztosított adathalmazok közül mely adattípusok alapján készítette el a közzétett jelentést, tehát ilyen összehasonlításoknál nem lehettünk volna biztosak abban, hogy ugyanazt a bemeneti adatot használjuk a jelentések generálásához, mint amelyikből a meglévő jelentést író személy dolgozott.

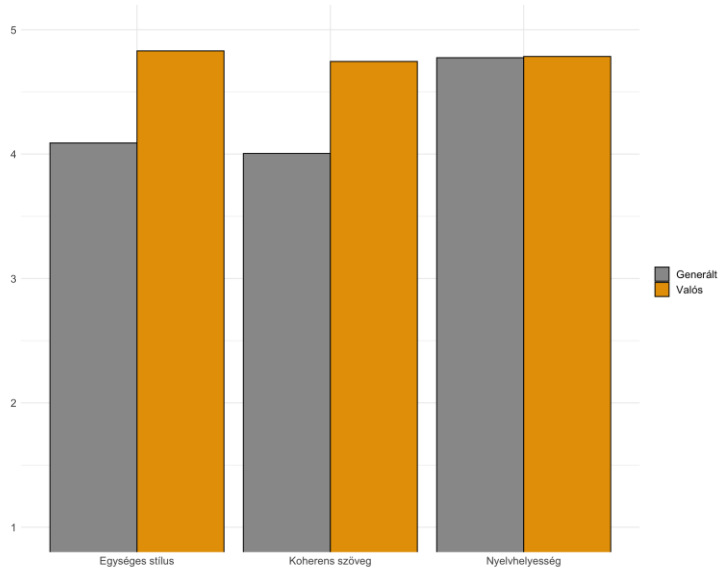

Fig. 2. a generált és a valós időjárásjelentésekhez tartozó minősítő szempontok átlagértékei

A jelentések értékelésére a projekten nem dolgozó személyeket kértünk fel, akik három szempont szerint minősítették az összesen 200 darab szöveget. A minősítés során az egységes stílust, a szöveg koherenciáját, illetve a szöveg nyelvhelyességét értékelték 1-től 5-ig terjedő skálán, ahol az 1-es érték a legrosszabb, az 5-ös pedig a legjobb minősítést jelentette. Minden időjárásjelentést egy annotátor értékelt.

Ahogyan a fenti ábrán látszik, az általunk generált jelentések – a nyelvhelyességet kivéve – nem érték el a valós időjárásjelentések minőségét, azonban nem is maradtak el azoktól jelentősen, és mindegyik értékelési szempont szerint meghaladták a 4,0-s átlagot.

4 Kihívások és limitációk, továbbfejlesztési lehetőségek

Az automatikus időjárásjelentés generálással kapcsolatos limitációk egy része az elérhető adatok körére vonatkozik. Jelentős limitációt jelent a mért adatokra vonatkozó fájlok adattartalma. Ahogy említettük, ezekben csupán hét település adatai szerepelnek, ami lehetetlenné teszi a régiós vagy megye szintű összehasonlításokat a mért értékek és az előrejelzések között. További problémát jelentenek az ezekben a fájlokban mutatkozó adathiányok az időjárás kódok esetében. Ezek Győr és Sármellék állomásoknál szinte teljesen hiányoznak a projektben felhasznált több mint fél éves adathalmaz alapján, Pogánynál pedig 21 százalékban mutatkozik adathiány. Mivel ezek a települések mindegyike a Dunántúlon található, a nagyrégiós összehasonlítások is rendkívül bizonytalanok lennének az adathiányok miatt. A fentiek miatt csak országos szintű összehasonlításokat tudtunk kalkulálni.

Ennél kisebb jelentőségű, de mindenképpen említésre méltó probléma, hogy a háromnapos előrejelzések óránkénti, a nyolcnapos előrejelzések pedig napi felbontásúak, azonban ahogy láttuk, ezek az állományok eltérő adatokat tartalmaznak. Ez azt jelenti, hogy a nyolcnapos előrejelzésekből származó, csapadékkal és széllel kapcsolatos kalkulációinkat csak napi szinte tudjuk kiszámítani, tehát itt nincsen lehetőségünk arra, hogy az egyes napszakokat megkülönböztessük. Így tehát az egyes napszakokra csak az időjárás kódok és a hőmérsékletértékek alapján tudunk előrejelzéseket adni. További korlátot jelent az elkészült időjárásjelentések információtartalmára vonatkozóan, hogy nem áll rendelkezésünkre olyan adat, amiből szélirányt vagy felhőzettípust tudunk előrejelezni, ezért ezek kimaradnak a program által generált jelentésekből. Ezen kívül szintén nincsenek olyan adataink, amelyek alapján például olyan, komplexebb kijelentéseket tehetünk, hogy „egy nyugatról érkező melegfront vonul át az ország felett, jelentős csapadékot hozva először a Dunántúlon, majd a középső országrészben".

Munkánk során arra törekedtünk, hogy az általunk megírt kódok könnyen bővíthetők legyenek további adattípusokkal (pl. további OMSZ adatokkal) és egyéb logikai műveletekkel. A kiterjesztés történhet például eltérő hosszúságú időszakokra (hétvégi vagy többnapos jelentések megírására), vagy különböző speciális tartalmú jelentésekre (pl. nyáron vizek hőmérséklete, télen hójelentés síterepekről, UV előrejelzés stb.)

A generált időjárásjelentések minőségének további javításához fontos irányt jelenthet a meteorológus szakértőkkel való közös munka, amelynek során az általunk meghatározott thresholdok és fontosság értékek tovább pontosítására nyílna lehetőség.

Szintén a további fejlesztések egy lehetséges irányát jelenti megfelelő méretű tanítóadatbázisok létrehozását követően az OpenAI-ban elérhető fine-tuning használata, amelynek során akár bizonyos újságokban szereplő időjárásjelentések speciális nyelvezetére is rátanítható lenne a nyelvi modell. Ez egyben az egyik lehetséges módja a generált szövegek további javításának. A másik utat a meglévő kódok magyar nyelvű nyelvi modellre való átültetése jelenti a jelenlegi angol helyett, amely az automatizált fordítás során megjelenő esetleges minőségromlást is kiküszöbölné. GPT-3 tudású magyar nyelvi modell (még) nem áll rendelkezésre, de GPT-2-re építve már voltak próbálkozások magyar nyelvű tartalmak előállítására (Yang Zijian, 2022).

Bibliográfia

Belz, A. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. Natural Language Engineering, 14(4). pp. 431–455. (2008)

Floridi, L., Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4). pp. 681–694. (2020)

Gatt, A., Krahmer, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61. pp. 65–170. (2018)

Goldberg, E., Driedger, N., Kittredge, R. I. Using natural-language processing to produce weather forecasts. IEEE Expert, 9(2). pp. 45–53. (1994)

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving language understanding by generative pre-training. (2018)

Ramos-Soto, A., Bugarin, A. J., Barro, S., Taboada, J. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. IEEE Transactions on Fuzzy Systems, 23(1). pp. 44–57. (2014)

Reiter, E., Sripada, S., Hunter, J., Yu, J., Davy, I. Choosing words in computer-generated weather forecasts. Artificial Intelligence, 167(1-2). pp. 137–169. (2005)

Turner, R., Sripada, S., Reiter, E., Davy, I. P. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In: International conference on innovative techniques and applications of artificial intelligence. pp. 75–88. Springer, London (December 2007)

Yang, Z. Gy. „Az invazív medvék nem tolerálják az agressziót". Magyar GPT2 kísérleti modell. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2022. január 27–28. pp. 464–476. Szegedi Tudományegyetem (2022)