Villámgyors magyar nyelvű érvelő modellt fejlesztettek az ELTE IK kutatói

Az ELTE Informatikai Kar Mesterséges Intelligencia Tanszékének kutatói a Digitális Örökség Nemzeti Laboratóriummal együttműködésben először tanítottak magyarra nagy érvelő nyelvi modellt.

A mesterséges intelligencia mérnökökből és nyelvtechnológusokból álló kilencfős csapat a költséghatékony tanítás során körülbelül 200 millió oldalnak megfelelő szöveget dolgozott fel a magyar akadémiai közösség számára elérhető Komondor HPC infrastruktúrán – hazánk legnagyobb szuperszámítógépén.

Munkájuk eredményeként a Racka-4B modell teljesítménye a magyar nyelvi feladatokban a kétszer akkora (8 milliárd paraméteres) modellek teljesítményével is összemérhetővé vált, sebessége pedig jócskán meghaladta azokét.

A szerzők tanulmányukért és bemutatott prezentációjukért megkapták a legjobb publikációnak járó díjat az idei, XXII. Magyar Számítógépes Nyelvészeti Konferencián (MSZNY).

Miért van szükség saját, magyar fejlesztésű nyelvi modellekre?

A globális technológiai óriások által fejlesztett nyelvi modellek, bár hatalmasak, a magyar nyelvvel és kultúrával még gyakran meggyűlik a bajuk. Ennek egyik oka, hogy a magyar morfológiailag egy rendkívül gazdag nyelv: szavaink sok ragot, jelet, képzőt hordoznak, így egyetlen szóalakban rengeteg információ sűrűsödhet össze. Ráadásul nincs olyan, a magyarhoz közeli rokon nagy világnyelv, amelynek digitális jelenléte – technológiai értelemben – magával húzná a magyart. Ha viszont ezek a rendszerek nem értik és beszélik elég jól a nyelvünket, nem integrálják a régió kultúráját és történeti tudását, vagy akár a hazai jogszabályokat, a jogi szaknyelvet, akkor Magyarország több téren is hátrányba kerülhet. Ahhoz, hogy a régió megőrizze digitális szuverenitását, elengedhetetlen a saját adatokon tanított technológia.

Erre a kihívásra válaszul született meg korábban a Puli modellcsalád, a technológia gyors fejlődésével azonban megjelent az igény az összetettebb, úgynevezett érvelő (reasoning) képességgel rendelkező rendszerek fejlesztésére is.

Ezt az űrt tölti be most a Racka. A modell egy nyílt forráskódú, Qwen3-4B alapokon nyugvó, úgynevezett paraméterhatékony (LoRA) eljárással magyarított rendszer, amely a korábbi modellekkel szemben logikai és érvelő képességekkel is rendelkezik.

Kihívások és technológiai válaszok a magyarítás során

A modell magyarítása több párhuzamos technológiai lépésben történt, melyek közül az egyik legfontosabb a mesterséges intelligencia „szótárának” optimalizálása volt. A nyelvi modellek a szövegeket feldolgozáskor apró egységekre, úgynevezett tokenekre bontják. A döntő arányban világnyelveken tanított nemzetközi modellek szótára azonban alapvetően az angol nyelvre van optimalizálva, ami azt eredményezi, hogy a magyar szavakat túl sok, apró, logikátlan darabra vágják szét.

A kutatócsoport azzal a mérnöki megoldással élt, hogy az eredeti modell mintegy 150 ezer elemből

álló szótárából eltávolítottak nagyjából 32 ezer olyan ritka tokent (például bizonyos távol-keleti karaktereket), amelyek a projekt szempontjából lényegtelenek voltak. Ezek helyére pedig kifejezetten a magyar nyelvre optimalizált tokeneket illesztettek be, és úgy hangolták át a rendszert, hogy előnyben részesítse ezek használatát. Ennek a bravúrnak köszönhetően a modell 47%-kal kevesebb tokenből tudja felépíteni ugyanazt a magyar szöveget, ami a gyakorlatban azt jelenti, hogy drasztikusan felgyorsult a szöveggenerálás, ráadásul a nyelvtani és ragozási hibák is jelentősen ritkultak. A modellt összesen 160 milliárd tokennyi adaton tanították tovább.

Bár a fókusz a magyar nyelven volt, az adathalmaznak csak a 44%-át tette ki a hazai szöveg, a maradék angol (24%), német (21%), illetve programkód (11%) volt. Az idegen nyelvű tanítóadatok használatának oka, hogy ha a modellt kizárólag magyar adattal bombázzák, felléphet az úgynevezett katasztrofális felejtés (catastrophic forgetting) jelensége, vagyis a rendszer elveszítheti a korábban már megtanult, értékes általános képességeit. Az angol és német nyelv – amelyek a statisztikák szerint a leggyakoribb idegen nyelvek hazánkban – biztosították a széles körű tudás megmaradását. A programkódok betáplálása pedig az alapmodell logikai és érvelési (reasoning) képességeinek megtartása miatt volt kulcsfontosságú.

Jövőbeli tervek egy teljes modellcsaláddal

A kutatócsoport célja a hazai tudományos szféra támogatása, így a Racka modell kutatási és fejlesztési célokra szabadon, nyíltan elérhető. Sikerét és hiánypótló mivoltát jól mutatja, hogy csak az elmúlt hónapban több mint 600 alkalommal töltötték le a projekt Hugging Face oldaláról.

Az akadémiai kutatás nélkülözhetetlen, de az ilyen költséges és erőforrás-igényes fejlesztések esetén kiemelten fontos, hogy a projekt valós felhasználási igényekhez kapcsolódjon. A Racka fejlesztésén dolgozó kutatócsoport azonban nem egyetlen modellben, hanem egy egész modellcsaládban gondolkodik, azt tervezve, hogy kilép a kelet-közép-európai regionális piacra is. Bár adatbiztonsági és elérhetőségi szempontból szükség van kifejezetten kis méretű, helyi szervereken (vagy akár mobiltelefonokon) biztonságosan futtatható modellekre is, ugyanakkor vannak olyan komplex feladatok – mint például a hosszú dokumentumok értelmezése vagy a bonyolult következtetések levonása –, amelyekhez egy sokkal nagyobb, általános tudással rendelkező rendszer kell. Ennek a nagyobb léptékű, regionális tudást is integráló modellnek az előkészítése és fejlesztése már zajlik, szoros együttműködésben a Mynds.ai piacorientált céggel. A projektet az újonnan kiépülő európai MI-infrastruktúrán és a barcelonai MareNostrum 5 szuperszámítógépen tervezik megvalósítani.

Sajtókapcsolat:

- Horváth Judit
- ELTE IK
- +36 30 224 0998
- horvathjudit@inf.elte.hu

Eredeti tartalom: Eötvös Loránd Tudományegyetem

Továbbította: Helló Sajtó! Üzleti Sajtószolgálat

Ez a sajtóközlemény a következő linken érhető el:
https://hellosajto.hu/?p=30949