

Önálló laboratórium beszámoló

Távközlési és Mesterséges Intelligencia Tanszék

|  |  |
| --- | --- |
| Készítette: | **Molnár Botond Kristóf** |
| Neptun-kód: | **H1XYPA** |
| Ágazat: | **Információs rendszerek** |
| E-mail cím: | **molnarboti2003@gmail.com** |
| Konzulens(ek): | **Dr. Mihajlik Péter** |
| E-mail címe(ik): | **mihajlik@tmit.bme.hu** |

Téma címe: Nagy nyelvi modellek és beszédfelismerési modellek integrálása

Feladat

A természetes beszéd automatikus szöveggé konvertálása még mindig jelentős kihívás, ha a beszélők szétválasztása, az írásjelezés, névelemek tag-elése, az idegen nyelvű kifejezések felismerés és a zajban is nagy pontosság a követelmények között van. Noha a klasszikus/neurális nyelvmodellek alkalmazása alapvetőnek számít ezen a területen, a (nagyon) nagy nyelvi modellek (LLM-ek, pl. GPT-4, ChatGPT, LLAMA, BARD, stb.) felhasználása egyáltalán nem magától értetődő. A hallgató feladat feltárni a közvetlen, beszéd-szöveg konverziót segítő és az utólagos (pl. korrekciós) LLM alkalmazásokat.

**2024/2025. 2. félév**

# A laboratóriumi munka környezetének ismertetése, a munka előzményei és kiindulási állapota

## 1.1 Bevezető

A féléves laboratóriumi feladatom célja az volt, hogy egy automatikus beszédfelismerő rendszer (Automatic Speech Recognition – ASR) kimenetét nyelvi modellek segítségével összefoglaljam. Az ASR rendszerek mélytanulási alapú modellek, amelyek képesek hangból természetes nyelvű szöveget előállítani. Ezek a kimenetek azonban gyakran nyersek, redundánsak vagy nehezen áttekinthetők, így különösen fontos, hogy a lényeget informatív és tömör módon összegezzük.

Az összegzéshez nagy nyelvi modelleket (Large Language Models – LLM-ek) használtam, amelyek szintén mélytanulási modellek. Ezek hatalmas szöveges adathalmazon tanulnak, és képesek a természetes nyelvű szövegek értelmezésére, a kontextus felismerésére, valamint új, nyelvtanilag helyes és koherens szövegek generálására. Kiválóan alkalmazhatók szövegösszegzésre, mivel képesek a tartalom lényegi elemeire fókuszálni.

A konzulensem javaslatára a llama-cpp keretrendszert alkalmaztam, mivel ez lehetővé teszi kisebb LLaMA modellek hatékony, lokális futtatását. A Meta által fejlesztett LLaMA (Large Language Model Meta AI) modellek nyíltan elérhetők, és széles körben támogatottak különböző kvantálási szinteken (pl. 4-bit, 5-bit, 8-bit), ami jelentős erőforrás-megtakarítást tesz lehetővé. A kvantált modellek előnye, hogy kisebb memória- és számítási igénnyel futtathatók, miközben megtartják a szükséges nyelvi teljesítményt. Emellett több előre finomhangolt változatot is kipróbáltam, amelyek specifikusabb feladatokra lettek optimalizálva, így lehetőségem volt összehasonlítani a különböző konfigurációk viselkedését az összegzési feladat szempontjából.

Ez azért volt lényeges szempont, mert a rendszer a későbbiekben várhatóan nagyszámú lekérdezést fog kezelni, így előnyös, ha a modellek gyorsan és alacsony erőforrásigénnyel futtathatók. A kiválasztott kisebb modellek megfelelő egyensúlyt kínálnak a futtathatóság és a szövegértési képességek között.

Ez a munka egy kari szinten zajló nagyobb projektbe illeszkedik, amelynek célja egy olyan intelligens rendszer kialakítása, amely képes folyamatos beszéd feldolgozására és annak strukturált visszacsatolására, ezzel támogatva a hatékonyabb ember–gép interakciót és információkinyerést.

## Elméleti összefoglaló

Amikor a tágabb tudományos vagy műszaki környezetről beszélünk, akkor azt értjük alatta, hogy „művelt laikus” – alapesetben ennek tekinthető a tárgyfelelős – megtalálja azokat a kapcsolódási pontokat, amelyek segítségével az ő ismereteihez a beszámolóban tárgyalt témakör és az elvégzett munka csatlakoztatható. Ez mindenképpen, szükséges, hiszen enélkül az olvasónak előzetes tájékozódást kellene végeznie a szűkebb szakterületen, hogy az elvégzett munka jellegét, súlyát, nehézségét meg tudja ítélni.

Természetesen a terjedelmi korlátok miatt nem lehet teljes mértékben bemutatni az adott szűkebb szakterületet, ezért meg kell adni az érdeklődő olvasónak a lehetőséget a további tájékozódásra. Részben erre szolgálnak a hivatkozások[[1]](#footnote-1), lábjegyzetek.

Nagyon fontos, hogy abban az esetben, amikor a hallgató féléves munkájának egy része vagy egésze az, hogy tanul, hogy maga is ismerkedik azzal a szakterülettel, amelyen dolgozni fog, akkor az tudatosan törekedjen arra, hogy az Elméleti összefoglaló és az Elvégzett munka ismertetése szétválasztható legyen. Az előbbinek a szakterület alapvető, általános ismereteit kell tartalmaznia, amelyekről a „művelt laikusnak” részleges ismeretei lehetnek. Míg az elvégzett munka leírásába azoknak a specifikus ismereteknek a bemutatása kerülhet, amelyek a később elvégzett vagy elvégzendő feladatokat konkrétan megalapozzák.

További tanulmányozásra ajánljuk Eco professzor művét [1].

A hivatkozások kezelésénél fontos, hogy mindig a mondat részeként tekintsünk rá, és ha szükséges, akár többször is hivatkozzunk meg egy forrást, de az első előfordulásakor mindenképpen. Ugyanez érvényes a rövidítésekre: minden rövidítést a legelső előforduláskor magyarázni kell, később viszont használhatók a rövidített formák is. Például: a Tiger Tree Hash (TTH) a hashelés egy speciális formája.

## 1.3 A munka állapota, készültségi foka a félév elején

A félév kezdetén a munka alapját képező automatikus beszédfelismerő (ASR) modell már rendelkezésre állt, azonban a szövegösszegző komponens kidolgozásával korábban még nem foglalkoztak. Az összegzés lehetősége eddig csupán ötletként merült fel, tényleges megvalósítás vagy tesztelés nem történt. Ennek megfelelően az összefoglalási feladatot teljes mértékben a nulláról kellett elindítani. Segítségként a konzulensem ajánlotta a llama-cpp keretrendszert és néhány kisebb LLaMA modell kipróbálását, de a munka szakmai és technikai kidolgozása alapvetően önállóan történt.

# Az elvégzett munka és eredmények ismertetése

## 1.1 <A munkám ismertetése logikus fejezetekre tagoltan>

<Én magam (nem a társam!!) a félév során következőket olvastam el / programoztam / készítettem el / teszteltem / dokumentáltam / néztem át / tanultam meg, stb. Tételes leírása és nem felsorolása mindannak, ami a félév során történt, alátámasztandó azon állításom a konzulens/tárgyfelelős felé, hogy összességében mindent beleértve tényleg dolgoztam a TVSZ szerint kreditenként 30 órát, azaz a heti 2 kontakt órás tárgy esetében min. 2,5\*30 = 75 munkaórát, illetve a heti 6 kontakt órás tárgy esetében min. 8\*30 = 240 munkaórát...>

Ebben a részben a hallgató az általa elvégzett munkát mutatja be. Hangsúlyosan a saját munka bemutatása a cél, hiszen a hallgató ezzel igazolja a témavezető és a tárgyfelelős irányába, hogy – folyamatosan fejlődve és egyre több és jobb munkát végezve – a szakdolgozatát/diplomadolgozatát képes lesz megírni. A beszámoló nem munkanapló, nem arra vagyunk kíváncsiak, hogy mit mikor csinált a hallgató és mennyi időt töltött vele, hanem egy eredmény-centrikus beszámolót szeretnénk olvasni. De itt is fontos tudni, hogy megosztott feladat esetén ki-mit csinált, mekkora részt vállalt.

Az egész beszámoló elkészítésénél törekedni kell a magyar nyelv szabályainak követésére és a műszaki dokumentáció/tudományos közlemény írásával kapcsolatosan kialakult közmegegyezés szerinti formai követelmények betartására. Tehát nem kell többes számként hivatkozni *saját magunkra,* kerülni kell a furcsa megfogalmazást, passzív és egyéb kifacsart mondatszerkezeteket. Az *egy* szót határozatlan névelőként történő használatakor ne írjuk ki számként!

A beszámoló természetesen nem csak szöveget tartalmazhat, hanem képleteket, táblázatokat, ábrákat és még sok minden mást. Ezek kapcsán az alábbi elvek irányadók:

* Az ábráknak, képeknek és táblázatoknak mindig van számuk és címük. A cím nem ennyi: 1. ábra, hanem azt írd le, ami látható rajta.
* Az ábrákra, a képekre és a táblázatokra a szövegben hivatkozni kell, és a szövegben elemezni kell azokat. Például az 1. ábrán látszik, hogy a vizsgált félévben még két napos csúszással is lehetett jeles érdemjegyet szerezni a tárgyból, de a pontosság még nem garancia a jó jegyre: öten nem kaptak jelest, noha nem késtek a leadással.
* Az ábrák, képek és táblázatok mérete a szükségesnek megfelelő legyen: elég nagy ahhoz, hogy kinyomtatva is olvasható és értelmezhető legyen, de nem nagyobb annál, mint amit szerepe indokol.
* A grafikonoknak a tengelyeken legyenek feliratai és ha releváns, a mértékegység is.
* A képletek esetében nem minden képletre történik hivatkozás, de ahol igen, ott a képletet a műszaki irodalomban jellemző módon a sor végére tett kerek zárójelben lévő számmal jelöljük meg. A képleteket ne képként illeszd be a szövegbe!
* Kódrészleteket, ha nem relevánsak, ne illeszd be képként, főleg ne rossz minőségben. Nyugodtan teheted függelékbe és hivatkozd be a szövegben, mint a képeket, pl. Az 1. számú függelékben található az adatbeolvasó kód, melyet C++ nyelven készítettem el.

|  |
| --- |
| 1. ábra. Hallgatók érdemjegyeinek eloszlása az írásbeli beszámoló késése függvényében |

Az írásbeli beszámolót a témavezető és a tárgyfelelős is értékeli. A tárgyfelelősi értékelés szempontjai az alábbiak:

1. Megfelel-e az elvégzett munka a félév elején kiadott feladatnak?
2. Megfelel-e a beszámoló a formai követelményeknek? Ezen belül:
   1. Megfelelő-e az elméleti bevezető és az irodalomjegyzék?
   2. Egyértelmű-e, hogy mi volt a hallgató saját munkája?
   3. Megfelelő-e a dokumentum technikai színvonala?

Ezen kívül a tárgyfelelős veszi figyelembe az értékelés során kialakult félévi jegyre vonatkoztatva az ún. „hanyagsági faktor” értékét, amelyet (1) szerint állapítunk meg.

|  |  |
| --- | --- |
|  | (1) |

Az (1)-ben szereplő a szám a munkaterv beadásában történt késedelemre, míg a b szám az írásbeli beszámoló beadásában történt késedelemre vonatkozik. Utóbbi értékeiről az 1. táblázat tájékoztat.

|  |  |
| --- | --- |
| Az írásbeli beszámoló beadásának napja  a szóbeli beszámolóhoz képest (munkanapban) | A „b” faktor értéke |
| **-4. nap** | **0.04** |
| **-3. nap** | **0.09** |
| **-2. nap** | **0.20** |
| **-1. nap** | **0.30** |

1. táblázat. Az írásbeli beszámoló késedelmes beadásával kapcsolatos hanyagsági faktor értéke

A beszámoló értékeléséről részletesebben írunk [3]-ban.

Itt ismét megemlítjük a rövidítéseket. Ezeket a rövidítéseket, betűszavakat néhány, az infokommunikáció területén nagyon ismert és gyakran használt kifejezéstől (például IP, TCP, GPRS, UMTS) eltekintve ki kell fejteni logikusan az *első használat* alkalmával (például így: „A GPS (Generalized Processor Sharing) egy ideális folyadékmodellen alapuló csomagütemező eljárás.”).

A beszámoló készítése során előfordulhat, hogy a hallgató úgy érzi, hogy alfejezetekkel tagolva jobban olvasható és érthető lenne a beszámoló. Ennek akadálya nincs, de érdemes arra figyelni, hogy a túlzott tagolás sem tesz jót egy írásműnek, illetve hogy a címsorokban a rövidítések és a hivatkozások használata tilos. Tartalomjegyzéket készíteni nem szükséges a beszámolóhoz, de nem is tilos, kivéve azt az esete, amikor nyilvánvalóan terjedelemnövelési célokat szolgál.

A beszámoló terjedelme tárgyanként változhat. Általános szabály, hogy 1 hüvelyknél nagyobb margókat ne használjunk, a szöveg legyen egyszeres sortávú, sorkizárt és 12 pontos betűméretű. A bekezdések kezdődjenek behúzással a minta szerint.

## 2.2 Összefoglalás

Ebben a részben az *adott* *félévre vonatkozó, az Önálló laboratórium tárgy keretében elvégzett munka során* elért ***új*** eredmények ismételt, vázlatos, **tömör** összefoglalását várjuk, lehetőleg nem felsorolásként. Itt még egyszer ki lehet térni a leglényegesebb eredményekre, valamint a félév során felmerülő nehézségekre, de meg lehet említeni a továbbfejlesztési irányokat, lehetőségeket is.

Ezt a részt tagolható a következő pontok megválaszolásával:

* Mi volt az aktuális kérdés, **probléma,** amivel a félév során foglalkoztál?
* Mi a dolgozat **célja**, miért érdekes egyáltalán ezzel a problémával foglalkozni?
* Milyen **módszereket** használtál a probléma megoldása érdekében?
* Mik a legfontosabb **eredmények**?
* Milyen **következtetéseket** lehet levonni?

Ha valaki elolvassa ezt a részt, képet kell kapnia az egész dolgozatról!

Fontos, hogy az itt megadott sablontól el lehet térni, használata nem kötelező, csak segítséget jelenthet, viszont a fedőlap lehetőleg maradjon ugyanez és tartalmilag egyezzen meg a sablon irányelveivel. A beszámoló felépítésében nem érdemes eltérni a Bevezető - Féléves munka és eredmények bemutatása - Összefoglaló hármastól.

# Irodalom, és csatlakozó dokumentumok jegyzéke

## A tanulmányozott irodalom jegyzéke:

|  |  |
| --- | --- |
| [] | Umberto Eco, *Hogyan írjunk szakdolgozatot?,* Kairosz Kiadó, 2000, ISBN: 9639137537. |
| [] | Esterházy Péter, *Termelési-regény (Kisssregény)*, Magvető Könyvkiadó, 2004, ISBN: 9631423948. |
| [] | *Tájékoztató a Műszaki Informatika Szak önálló laboratórium tantárgyainak 2008/9. tanév I. félévi lezárásáról a BME TMIT-en* (VITMA367, VITMA380, VITT4353, VITT4330), http://inflab.tmit.bme.hu/08o/lezar.shtml, szerk.: Németh Felicián, 2008. november 5. Utolsó letöltés ideje: 2010-10-12 |

A tanulmányozott irodalmat hivatkozni kell a szövegben! Szükség esetén többször is! Az irodalomjegyzék célja ugyanis kettős:

1. Az olvasó tájékoztatása, hogy a dokumentumban ki nem fejtett dolgoknak, a tudottnak vélt ismereteknek hol lehet bővebben utánanézni. Következésképpen ott kell meghivatkozni az irodalmat, ahová az irodalom kapcsolódik.

2. Megmutatni a tárgyfelelősnek/konzulensnek az elolvasott irodalom mennyiségét.

Javasoljuk, hogy a hallgatók tanulmányozzák, hogyan néznek ki a hivatkozások a villamosmérnöki/informatikai szakma vezető szakmai folyóirataiban megjelenő cikkekben. Ebben a témavezető is biztosan tud segíteni.

A hivatkozás teljességére és egyértelműségére tessék ügyelni! Például, ha egy könyvnek több, eltérő kiadása is van, akkor azt is meg kell jelölni, hogy melyik kiadásra hivatkozunk.

A webes hivatkozások problémásak szoktak lenni, de manapság egyre több az olyan dokumentum, ami csak weben lelhető fel, ezért használatuk nem zárható ki. Itt is törekedni kell azonban a pontosságra és a visszakereshetőségre. A weben található dokumentumoknak is van címe, szerzője, illetve meg kell adni a letöltés/olvasás időpontját is, hiszen ezek a dokumentumok idővel megváltozhatnak.

A wikipédiás hivatkozások használata nem javasolt.

Nem publikus dokumentumok hivatkozása nem javasolt és csak kivételes helyzetben elfogadható!

## Csatlakozó egyéb elkészült dokumentációk / fájlok / stb. jegyzéke:

<A munka ezen beszámolóba be nem fért eredményeinek (pl. forrás fájlok, mindenképpen csatolni akart forráskód részlet, felhasználói leírások, programozói leírások (API), stb.) megnevezése, fellelhetőségi helyének pontos definíciója, mely alapján a az erőforrás előkereshető – értelemszerűen nem nyilvános dokumentumok hivatkozása nem elfogadható.>

1. A másik nagyon fontos céljuk az állítások alátámasztása. A lábjegyzetek használatát egyébként nem érdemes túlzásba vinni, mert állandóan megtörik az olvasás folyamatát (lásd például [2]). Akkor kell használni, amikor a lábjegyzetben közlendő információ érdekes lehet, de nem tartozik közvetlenül a tárgyhoz. Mindenképpen kerülendő az irodalmi hivatkozások lábjegyzetben való megadása. [↑](#footnote-ref-1)