Jobban fordítunk magyarra, mint a Google!

Laki László János, Yang Zijian Győző

Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
{laki.laszlo, yang.zijian.gyozo}@nytud.hu
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/a.
{laki.laszlo, yang.zijian.gyozo}@itk.ppke.hu

Kivonat Kutatásunkban célul tűztük ki, hogy áttekintjük az elérhető gépifordító architektúrákat és megvizsgáljuk, hogyan teljesítenek angol-magyar környezetben. A magyar, mint célnyelv nagy kihívás a fordítórendszerek számára, mivel mind grammatikai, mind szórendi szempontból eltért az angoltól. Kipróbáltunk ipari és kutatásra szánt gépifordító rendszereket egyaránt. Kutatásunk legfontosabb eredménye, hogy az általunk létrehozott modellek (MarianNMT és BART) nagymértékben jobban teljesítettek, mint a piacvezető multinacionális cégek rendszerei. Végül sikerrel finomhangoltunk (fine-tuning) angol-magyar fordításra egy többnyelvű T5 nyelvmodellt, aminek az eredménye összehasonlítható volt az elérhető modellekével.
Kulcsszavak: neurális gépi fordítás, Marian-NMT, BART, mT5

1. Bevezetés

Az Infokommunikációs és Információtechnológiai Nemzeti Laboratóriumot (InfoLab) létrehozó konzorcium tagjai (a Nemzetbiztonsági Szakszolgálat és az IdomSoft Zrt.) kettős célt tűztek ki maguk elé: kutatásaikkal támogatni szeretnék a feltörekvő infokommunikációs és információs technológiák biztonságos bevezetését és alkalmazását, valamint az államigazgatás digitális átalakulását.

Az IdomSoft Zrt. egyik kiemelt célja a Mesterséges Intelligencián (MI) alapuló technológiák közigazgatási célú alkalmazási lehetőségeinek kutatása és alkalmazása, mely lehetővé teszi, hogy az ügyfelek mentesüljenek minden olyan adat szolgáltatása alól, amely adat a közigazgatásban már rendelkezésre áll. A fejlesztések megkímélik az ügyfeleket mindazon szervezési és ügyintézési feladatoktól, amelyek a közigazgatási szervek közötti, belső igazgatásszervezéssel megoldhatóak. Cél emellett az érintésmentes, teljes mértékben digitalizált és automatizmusokkal segített, biztonságos és zökkenőmentes ügyintézés megteremtése is.

Ennek a stratégiai innovációnak részét képezi többek között a magyar nyelv sajátosságait készség szinten kezelő, a XXI. század elvárásainak megfelelő közigazgatási szolgáltatások megvalósíthatósága. Az IdomSoft Zrt. e célkitűzések

megvalósításában együttműködik hazai egyetemekkel, hogy azok K+F folyamatban megvalósított termékeit az általa fejlesztett közigazgatási informatikai megoldások kapcsán, a gyakorlatban alkalmazza.

A közigazgatás egyik fontos feladata az idegen nyelvű dokumentumok fordítása és a szakfordítók támogatása. Napjainkra a gépi fordítás napi szintű használati eszközzé vált. Az utóbbi években a neurális módszerek, azon belül is a transzformer modellek érték el a legjobb eredményeket a legtöbb nyelvtechnológiai feladatban, beleértve a gépi fordítást is. A neurális gépi fordítás eszközei látványosan jobb, emberek számára olvashatóbb fordítást produkálnak elődjeikhez képest, ennek köszönhetően alkalmassá váltak arra, hogy az emberi fordítók számára előfordítóként használják őket. A nemzetközi publikációk kiemelt figyelmet szentelnek az angol nyelvnek, és különböző architektúrákkal kísérleteznek egyre jobb fordítási teljesítmény elérése érdekében. Kutatásunk során ezeket a különböző modelleket vizsgáljuk fókuszba a magyart, mint célnyelvet állítva. Arra keressük a választ, hogy vajon miben különböznek ezek a módszerek egymástól, mennyire tudnak magyarra fordítani, illetve kvantitatív módszerekkel állapítjuk meg a közöttük fennálló különbséget. Modelljeink és szkriptjeink megtalálhatóak a Github[1] és Hugging Face[2] oldalainkon.

2. Neurális gépifordító rendszerek és módszerek

A cikkünkben összehasonlítottunk kutatásra szánt és ipari gépifordító rendszereket egyaránt. Az ipari termékek közül nem mindegyik rendszernél sikerült felderíteni a modell architektúráját és paramétereit. Ebben a fejezetben részletesen bemutatjuk a vizsgált rendszereket. A felsorolt lista tekinthető irodalomkutatásnak is, mivel szinte teljesen lefedjük az elérhető magyar nyelvű fordítórendszereket.

2.1. Marian NMT

A Marian NMT (Junczys-Dowmunt és mtsai, 2018) nevű keretrendszer egy C++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően[3] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019). A Marian NMT egy figyelmi (attention) modellel támogatott enkóder-dekóder architektúrájú neurális gépifordító modell. Legnagyobb előnye a többi módszerhez képest, hogy előtanított nyelvmodellek használata nélkül is a leggyorsabb futási idejű tanítást eredményezi. Két méretű transzformer modell tanítható be:

- Marian small: 6 réteg enkóder és 6 réteg dekóder; 8 figyelmi fej; 512 szóbeágyazás dimenzió; bementi hossz: 512; előre csatolt háló méret: 2048
- Marian big: 6 réteg enkóder és 6 réteg dekóder; 16 figyelmi fej; 1024 szóbeágyazás dimenzió; 1024 bemeneti hossz; előre csatolt háló méret: 4096

[1] https://github.com/nytud/machine-translation
[2] https://huggingface.co/NYTK
[3] https://marian-nmt.github.io/

2.2. BART és mBART

A BART (Lewis és mtsai, 2020) modell egy enkóder-dekóder architektúrán alapuló transzformer modell, amelyet a Fairseq (Facebook AI Research Sequence-to-Sequence Toolkit) fejlesztett[4]. Az enkóder kétirányú (Bidirectional), a dekóder autoregresszív (Autoregressive). A BART gyakorlatilag ötvöz egy BERT (Devlin és mtsai, 2019) és egy GPT (Radford és Narasimhan, 2018) típusú modellt. A kutatások alapján a BERT típusú modellek kiválóan alkalmasak szószintű és mondatszintű osztályozási feladatokra, azonban szöveggenerálás esetében gyengén teljesítenek. Ezzel ellentétben a GPT típusú autoregresszív modellek elsősorban szövegek generálására működnek jól, mint például szövegösszegzés vagy szöveggenerálás. A BART a két architektúra előnyeit ötvözi, ezért kiválóan alkalmas gépi fordításra. Jelenleg két különböző méretű BART érhető el: 1.) BART-base: 6 réteg enkóder és 6 réteg dekóder; 12 figyelmi fej; 768 szóbeágyazás dimenzió; bementi hossz: 512; 140 millió paraméter 2.) BART-large: 12 réteg enkóder és 12 réteg dekóder; 16 figyelmi fej; 1024 szóbeágyazás dimenzió; 1024 bemeneti hossz; 400 millió paraméter

Az mBART (Liu és mtsai, 2020) egy több nyelven előtanított, zajtalanító (denoising) autoenkóder modell, mely a seq2seq koncepción alapul. A többnyelvű zajtalanító előtanítás alkalmazható mind a felügyelt (supervised), mind pedig a felügyelet nélküli (unsupervised) gépi tanulás teljesítményének javítására. Az mBART felépítését tekintve a BART sémáját követi. A modell szerzői az előtanítás során fókuszálnak a többnyelvűségre, a betanított modell ezt követően kutatásukban kétnyelvű beállításban kerül finomhangolásra. Az előtanításhoz a Common Crawl adatbázisból kivonatolt 25 nyelvet tartalmazó CC25 (Wenzek és mtsai, 2020; Conneau és mtsai, 2020) korpuszt használták. Mind mondatszintű, mind pedig dokumentumszintű gépi fordításra alkalmazták a több nyelven előtanított modellt. Kiemelkedő jelentőségű, hogy kizárólag a seq2seq koncepció használatával tudták javítani a dokumentumszintű gépi fordítás minőségét, ez korábbi hasonló munkákhoz (Miculicich és mtsai, 2018; Li és mtsai, 2019) képest jelentős előrelépést jelent. Az mBART modellel végzett munka rámutat a többnyelvű előtanításban rejlő lehetőségek transzfer tanulásos (transfer learning) irányban való felhasználhatóságára.

Az mBART nem tartalmazza a magyar nyelvi tudást, ezért erre a célra saját angol-magyar nyelvű BART modellt tanítottunk.

2.3. T5 és mT5

A **T5** (Text-To-Text Transfer Transformer) (Raffel és mtsai, 2020), a Google kutatócsapata által publikált, modell és keretrendszer újfajta lehetőségeket kínál a nyelvfeldolgozási feladatok terén. A természetes nyelvfeldolgozás eszköztárában kiemelt szereppel bír a transzfer tanulás (transfer learning), melynek során a nyelvi modell egy adatokban gazdag feladaton van tanítva, majd ezt követően kerül finomhangolásra egy soron következő célfeladatra. Ideális esetben a modell

[4] https://github.com/pytorch/fairseq/tree/master/examples/bart

az előtanítás folyamán olyan általános tudásra tesz szert, amely átvihető és sikeresen alkalmazható a célfeladatok esetében is. A T5 projekt a transzfer tanulási alapelveket alkalmazza a szövegből szöveg (text-to-text) problémamegközelítés kontextusában. Kiindulási ötletként az szolgált, hogy minden szövegelemzési feladatot (fordítás, kérdések megválaszolása, osztályozás) szövegből szöveg problémaként közelít meg, azaz szöveg a bemenet és egy újabb szöveg lesz a kimenet.A szövegből szöveg eljárások nagy előnye a széleskörű alkalmazhatóság, ugyanis gyakorlatilag bármilyen természetes nyelvelemzési feladatra felhasználhatók, így például gépi fordításra, összefoglaló készítésére, kérdések megválaszolására vagy épp szentiment analízisre.

Az ilyen nagy volumenű kísérletekhez speciális korpusz szükséges, ennek érdekében lett létrehozva az ún. Colossal Clean Crawled Corpus (rövidítve C4), amely egy több száz gigabájtnyi világhálóról összegyűjtött és tisztított angol nyelvű szöveget tartalmaz. A C4 korpusz alapját a Common Crawl [5] adatbázis képezi. A transzfer tanulásos módszereknek fontos jellegzetessége, hogy az előtanításhoz jelöletlen adathalmazra van szükség. További kívánalmak egy ilyen keretrendszerben alkalmazható korpusz felé, hogy nagy méretű, változatos és magas minőségű legyen. Összehasonlításképpen, az alkalmazott C4 korpusz kétszer akkora, mint a Wikipédia, vagyis lényegesen nagyobb mennyiségű adatot biztosít. Az T5 esetében a paraméterek száma alapján 5 különböző méretű modell került kialakításra, ezek a következők: Small (300 millió paraméter), Base (580 millió paraméter), Large (1,2 milliárd paraméter), XL (3,7 milliárd paraméter), XXL (13 milliárd paraméter)

A T5 projekt eredményeként létrehozott keretrendszer kiváló eredmények elérését tette lehetővé. A 11 milliárd paraméterrel futatott legnagyobb csúcsteljesítményt nyújtott több tekintetben is, így például GLUE, SuperGLUE, SQuAD és a CNN/Daily Mail referencia feladatokban.

Az **mT5** (Xue és mtsai, 2021) a korábban tárgyalt T5 több nyelvre kiterjesztett verziója. Az mT5 létrehozása során a szerzők törekedtek arra, hogy minél inkább megőrizzék a kísérletek során több ízben is kiemelkedően teljesítő T5 strukturális jegyeit, ennek megfelelően az mT5 örökölte a szövegből szöveg (text-to-text) alapú problémamegközelítést és az általános előtanítás menetét is, amelyhez szintén rendkívül nagy méretű korpuszt használtak.

Az mT5 betanításához az mC4 korpuszt használták, amely a T5 tanítására alkalmazott C4 többnyelvű változata, és 101 különböző nyelvből szerepelnek benne szövegek. Többnyelvű modellek esetében általánosan felmerülő probléma, hogy amennyiben egy nyelv kevesebb forrással rendelkezik, akkor előfordulhat, hogy a gyakoribb mintavétel miatt a modell illesztése nem megfelelően alakul. Ezen probléma kiküszöbölésére korábbi modellek esetében is alkalmazott gyakoriságalapú mintavételi eljárást (Devlin és mtsai, 2019; Aharoni és mtsai, 2019) használtak a szerzők. Mivel az mT5 modell több, mint száz nyelvből álló korpuszon lett betanítva, ezért szükséges volt egy nagyobb méretű –250 ezer szóelemből álló –szótár alkalmazása is.

[5] https://commoncrawl.org

Az mT5 teljesítményének kiértékeléséhez az XTREME többnyelvű referenciarendszer (Hu és mtsai, 2020) 6 feladata lett alkalmazva. A keretrendszerben szerepelnek mondatpár, névelemfelismerés és kérdés megválaszolás feladatok is. Az eredmények tekintetében a legnagyobb modell, az mT5-XXL csúcsteljesítményt nyújtott kérdés megválaszolásos feladatokban.

Az mT5 projekt megmutatta, hogy a T5 modellrendszer kiválóan alkalmazható többnyelvű kontextusban is, és rendkívül erős eredményeket tud elérni különböző referenciafeladatokban. Az mT5 tartalmazza a magyar nyelvi tudást is, ezért kutatásunkban alkalmaztuk az mT5 small modellt gépi fordításra.

2.4. M2M100

Az M2M100 (Aharoni és mtsai, 2019) a Fairseq többnyelvű gépi fordítás projektje[6]. A többnyelvű gépi fordítás (multilingual machine translation) célja egy olyan átfogó modell megalkotása, amely bármely nyelvről bármely nyelvre képes fordítani. A gépi fordítás sokáig meglehetősen angolközpontúnak számított, azaz zömében olyan nyelvi modellek születtek, amelyek angolról vagy angolra fordítottak. A valóságban azonban a fordítás nem ilyen kizárólagos módon van felhasználva, tehát sok egyéb más nyelvről és más nyelvekre történő fordításra is komoly igény mutatkozik. A M2M100 projekt keretében 100 nyelvre készült el egy olyan adathalmaz, amely egy nagyfokú diverzitást hozó, az angol nyelv gépi fordítás tekintetében elfoglalt egyeduralmát megtörő módszertani újítás felé nyitja meg az utat. Ez végeredményben lehetővé teszi az újgenerációs többnyelvű gépifordító modellek megalkotását. A sok nyelvről sok nyelvre történő gépi fordításhoz nagy méretű adathalmaz létrehozására volt szükség. Az ilyen nagy volumenű soknyelvű adathalmazok generálásához szükséges az adatbányászat (Artetxe és Schwenk, 2019) és a visszafordítás (Sennrich és mtsai, 2016) alkalmazása. Az M2M100 tartalmazza a magyar nyelvi tudást, ezért kutatásunkban megmértük a teljesítményét.

2.5. Helsinki Marian NMT

A HNMT (Helsinki Neural Machine Translation) (Tiedemann és Thottingal, 2020) egy Marian NMT (Junczys-Dowmunt és mtsai, 2018) (small) rendszer, amely jelenlegi legjobban teljesítő fordító eljárás angolról finnre, ezen nyelvpár esetén a legmagasabb BLEU értékeket képes elérni. A HNMT mögött álló kutatók tesztelték a fordító rendszert angol-lett, angol-kínai és kínai-angol irányokban is, azonban ezeknél a nyelvpároknál szerényebb eredményeket értek el. A tanított gépifordító rendszer kifejezetten jól teljesít morfológiailag gazdag nyelvekre, például a finn nyelvre is.

A Helsinki Egyetem munkatársai céljai között volt, hogy lehetőleg a legtöbb nyelvre készítsenek gépifordító modelleket. Angol-magyar nyelvpárra több small modelljük is létezik. Ezeket teszteltük kutatásunkban.

[6] https://github.com/pytorch/fairseq/tree/main/examples/m2m_100

2.6. DeepL Fordító

A DeepL Fordító[7] egy ingyenes elérhető internetes fordító rendszer (DeepL GmbH, Köln, Németország). A fordító rendszer mögött álló vállalatot 2009-ben alapították Linguee néven, és az azonos névvel ellátott fordításokra specializálódott keresőmotort dobták piacra (deepl.com). A DeepL Fordító konvolúciós neurális hálókat (Kim, 2014) használ, és architektúrájának köszönhetően sokszor természetesebben hangzó fordításokat tud produkálni a piacon elérhető versenytársak megoldásaihoz képest. A 2018-ban elindított DeepL Pro egy tovább optimalizált verzióként lehetővé teszi, hogy a vállalat által fejlesztett mesterséges intelligencián alapuló megoldások még magasabb minőségű gépi fordításra legyenek képesek. 2021-ben 13 újabb európai nyelvvel, köztük a magyarral bővült a DeepL repertoárja.

2.7. Google Fordító

A Google Fordító[8] (Wu és mtsai, 2016) 2003-ban indult, akkor még statisztikai gépi fordítás elvén alapulva, majd 2016-ban ezt felváltotta a neurális háló alapú gépi fordítás. A neurális hálón alapuló megközelítés bevezetése lényegesen javított a fordítási minőségen, mivel szélesebb kontextus alapján következteti ki a jobban illeszkedő és ezáltal hitelesebb lefordított verziót[9]. A Google Fordító többféle lefordított verziót is listáz, így például az angolban nem specifikált, de franciában vagy spanyolban nőnemű és hímnemű megkülönböztetéssel létező szavak esetén először a nőnemű, majd a hímnemű verziót mutatja (Rescigno és mtsai, 2020). A Google Fordító 109 különböző nyelvet képes kezelni[10]. 2020-tól kezdve a szóban elhangzott szövegek lefordítása is lehetséges[11].

2.8. Yandex Fordító

A Yandex[12] egy orosz technológiai vállalat, amely gépi fordításon alapuló szolgáltatásokat értékesít a digitális termékek piacán. A Yandex által kifejlesztett fordító[13] két önálló gépifordító rendszeren alapul[14]. Az egyik egy statisztikai gépifordító, amely több százezer ugyanazon információkat tartalmazó, de különböző nyelveken íródott szövegek statisztikai összehasonlítása révén tanul. A Yandex statisztikai fordító mögött egy három komponensű gépifordító rendszer áll, melyek a fordító modell, a nyelvi modell és a dekóder. Magát a tényleges

[7] https://www.deepl.com/translator
[8] https://translate.google.com
[9] https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate
[10] https://translate.google.com/intl/en/about/languages
[11] https://blog.google/products/translate/transcribe-speech
[12] https://yandex.com
[13] https://translate.yandex.com
[14] https://yandex.com/dev/translate/doc/dg/concepts/how-works-machine-translation.html

fordítási folyamatot a dekóder végzi. Ennek során a fordító modell által felkínált különböző fordítási verziókat kombinálja össze és sorrendbe állítja az előfordulási gyakoriság alapján, amely a nyelvi modell révén kerül megállapításra. A másik fő komponens egy enkóder-dekóder architektúrájú neurális gépifordító. Tudomásunk szerint RNN architektúrájú (Cho és mtsai, 2014). A rendszer a két fordító rendszer által kiadott fordítást a CatBoost algoritmus (Prokhorenkova és mtsai, 2018) segítségével hasonlítja össze, majd a jobb fordítást adja végső kimeneteként.

2.9. Bing Fordító - Microsoft fordító

A Bing Fordító a Micorsoft Cognitive Services termékcsalád része, amely szövegek fordítását teszi lehetővé több, mint 100 különböző nyelven[15]. 2021-től már teljes dokumentumok lefordítására is használható. Kezdetben statisztikai megközelítést alkalmaztak a fejlesztők, majd 2018-ban átváltottak neurális háló alapú gépi fordításra. A Microsoft részéről is komoly érdeklődés övezi a többnyelvű gépi fordítást és számos kutatást végeznek a hatékonyság és a pontosság növelése érdekében. A többnyelvű és az egynyelvű modellek közötti fordítási pontosság tekintetében mutatkozó különbség leküzdésére Xu Tan és mtsai egy tudás desztilláción (knowledge distillation) (Bucila és mtsai, 2006) alapuló módszert dolgoztak ki (Tan és mtsai, 2019). A tudás desztilláció eredetileg modellek karcsúsításásra szolgáló módszer, amelynek keretében egy ún. diák modell kerül kialakításra, amely képes a tanár modell vagy több modell együttesének teljesítményét (adott esetben pontosságát) hozni. Ennek mintájára, egy-egy nyelvpárra szakosodott ún. tanár modellek tanítják be a diák modellt, amely az összes nyelvpárt egyetlen modellben kezeli. Két különböző eljárást fejlesztettek ki, az egyik a szelektív desztilláció, melynek során a desztilláció alkalmazását teljesítményalapú küszöbértékhez kötik, a másik pedig a Top-K desztilláció, amely a tanár modellek valószínűségi eloszlását vizsgálja és csak a legjobb együtthatóval rendelkező modelleket tölti be a memóriába. A módszer eredményességét tanúsítja, hogy a TED előadások átiratát tartalmazó szövegbankon tesztelve 44 nyelvről angolra fordítva, az összes nyelvre nézve 1 körüli vagy annál magasabb BLEU érték javulást hozott a kifejlesztett módszer alkalmazása.

2.10. eTranslation

Az eTranslation[16] egy automatizált fordítóeszköz, amellyel szövegrészleteket vagy akár teljes dokumentumokat lehet lefordítani az Európai Unió tagállamaiban használatos hivatalos nyelvekre, valamint izlandi, norvég, orosz, illetve egyszerűsített kínai nyelvre is. Az Európai Bizottság által rendelkezésre bocsátott gépifordító eljárás segítséget kíván nyújtani az Európai Unió kis- és középvállalkozásainak, közszolgáltatóinak, hivatalnokainak a gördülékeny kommunikáció és

[15] https://www.microsoft.com/en-us/translator/blog/2021/10/11/translator-now-translates-more-than-100-languages

[16] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

ügyintézés elősegítése érdekében. Az eTranslation könnyen integrálható más digitális rendszerekbe, amennyiben fordítási kapacitásra van szükség. A gépi fordítás folyamatának megkönnyítése érdekében számos előfeldolgozási és szövegtisztítási lépés is elérhetővé vált a CEF eTranslation Building Block projekt keretében. Példaként említhető, hogy hosszú mondatok könnyebb kezelhetősége érdekében ezek a fordítás előtt kisebb részekre vannak osztva, majd a fordítást követően koherens szöveggé kerülnek összeállításra. Az eTranslation rendszer nem általános, hanem specifikusabb szövegeken lett tanítva, mint például pályázati dokumentumok, jogi és orvosi szövegek stb. A modell több mint 1 milliárd mondaton tanult 24 nyelvre.

3. Mérések

3.1. Felhasznált korpusz

A gépi fordításhoz magunk építettünk egy angol-magyar párhuzamos korpuszt. A korpusz építéséhez az OPUS (Tiedemann, 2012) korpuszból vettünk angol-magyar (en-hu) párhuzamos alkorpuszokat. Felhasznált alkorpuszok: ParaCrawl, OpenSubtitles, Tatoeba, WikiMatrix, EUbookshop, PHP manual, TED2020, KDEdoc, KDE4. A korpusz méretei a 1. táblázatban láthatóak (nem tokenizált szövegen):

	Szegmens	Token		Type		Token átlag / mondat	
		en	hu	en	hu	en	hu
OpenSubtitles	42 655 519	272 571 665	209 481 645	2 382 239	6 519 406	6,39	4,91
ParaCrawl	12 681 746	196 278 983	172 671 171	3 555 484	5 713 776	15,48	13,62
WikiMatrix	488 319	8 978 943	7 673 323	627 814	1 057 487	18,38	15,71
TED2020	308 341	5 194 871	3 982 056	158 210	495 452	16,85	12,91
EUbookshop	438 264	9 406 548	7 847 111	360 311	648 778	21,46	17,90
KDE4	12 0657	622 959	649 457	62 257	98 940	5,16	5,38
Tatoeba	109 041	639 834	505 838	30 759	84 570	5,86	4,64
PHP	35 423	169 610	157 583	17 215	25 854	4,79	4,45
KDEdoc	861	10 904	9 474	2 402	2 987	12,66	11,00
Összesen	56 838 171	493 874 317	402 977 658	5 873 336	11 770 992	8,69	7,09

1. táblázat. Felhasznált részkorpuszok méretei.

3.2. Saját tanított gépi fordító modellek

Kutatásunk során kettő Marian, egy BART és egy mT5 gépi fordító modellt tanítottunk angol-magyar nyelvre.

A **BART** kísérlet esetében elsőként előtanítottunk egy angol-magyar kétnyelvű BART base modellt, ami elérhető a Hugging Face oldalon[17]. Az előtanításhoz az angol WikiText-103 (Merity és mtsai, 2017) és a Webcorpus 2.0 (Nemeskey, 2020) magyar Wikipédia részét használtuk. Az eredeti BART kutatáshoz hasonlóan csak azokat a bekezdéseket hagytuk meg, amelyek legalább egy darab pont írásjellel rendelkeztek. A felhasznált korpuszok méretei az 2. táblázatban láthatók (tokenizált szövegen).

	Angol WikiText-103	Magyar Wikipédia
Szegmensek száma	707.391	1.098.156
Tokenek száma	96.534.563	90.349.849
Type	596.820	3.137.980
Átlagos mondatszám / bekezdés	5	4
Átlagos tokenszám / bekezdés	125	69

2. táblázat. BART előtanításhoz használt korpuszok méretei.

Az angol-magyar BART-base modellünk előtanításához 4 darab GeForce GTX 1080 Ti (12GB) videokártyát használtunk, az alábbi paraméterekkel: batch méret / GPU: 12; szótár méret: 40.000; tanulási ráta: 2e-8; tanítási lépésszám: 170.000. Az előtanításhoz a Hugging face Transformers könyvtárban található Seq2SeqTrainer[18] és BartForCausalLM[19] függvényeket használtuk.

Az így előtanított BART modellünket tovább finomhangoltuk angol-magyar gépi fordítás feladatára. A finomhangoláshoz 4 darab GeForce GTX 1080 (12GB) videokártyát használtunk az alábbi paraméterekkel: batch méret / GPU: 26; maximum szöveghossz (bemeneti és kimeneti): 128; warmup: 15.000; fp16; epoch: 10; tanulási ráta: 5e-5. A finomhangoláshoz a Hugging Face Transformers könyvtárban található példakódot[20] használtuk fel.

A **Marian NMT** esetén a keretrendszer által alapértelmezetten biztosított paraméter beállításokat alkalmaztuk[21]. Az első esetben egy small modellnek megfelelő beállítást választottunk (Marian small), másodjára pedig egy kétszer több paraméterrel rendelkezőt (Marian big), amely elérhető a Hugging Face oldalon[22]. Subword tokenizáláshoz a beépített Sentence Piece (Kudo és Richardson, 2018) tokenizálót használtuk. A szótér mérete: 32.000.

Az **mT5** kutatásunkban az előtanított mT5 small[23] modellt finomhangoltuk angol-magyar nyelvre, amely elérhető a Hugging Face oldalon[24]. A tanításhoz 4

[17] https://huggingface.co/NYTK/translation-bart-128-en-hu
[18] https://huggingface.co/transformers/main_classes/trainer.html#seq2seqtrainer
[19] https://huggingface.co/transformers/model_doc/bart.html#bartforcausallm
[20] https://github.com/huggingface/transformers/tree/master/examples/pytorch/translation
[21] https://github.com/marian-nmt/marian-dev/blob/master/src/common/aliases.cpp
[22] https://huggingface.co/NYTK/translation-marianmt-en-hu
[23] https://huggingface.co/google/mt5-small
[24] https://huggingface.co/NYTK/translation-mt5-small-128-en-hu

darab GeForce GTX 1080 (12GB) videókártyát használtunk az alábbi paraméterekkel: batch méret: 6; prefix: „translate English to Hungarian: ”; maximum szöveghossz (bemeneti és kimeneti): 128; epoch: 1; tanulási ráta: 5e-5. Sajnos az epoch szám csak 1, amellyel közel egy hónap volt a futási idő. A finomhangoláshoz ugyanazt a könyvtárat használtuk, mint a BART finomhangolásnál.

3.3. Kipróbált gépifordítók és modellek

Kutatásunkhoz kipróbáltunk különböző kutatásra és ipari alkalmazásra szánt gépifordító rendszereket és modelleket, amelyek képesek angolról magyarra fordítani. Kísérleteinkben az alábbi rendszereket, modelleket próbáltuk ki:

- **M2M100**: Két modelljük is elérhető, egy kisebb (418M) és egy nagyobb (1.2B) modell. A fordításhoz a Hugging Face Transformers könyvtár M2M100Tokenizer és M2M100ForConditionalGeneration függvényeit használtuk.
- **Helsinki Marian NMT**: több modelljük is tartalmaz angol-magyar tudást: angol-magyar (en-hu); angol-finnugor (en-fiu); angol-uráli (en-urj); angol-multi (310 nyelv) (en-multi). A fordításhoz a Hugging Face Transformers könyvtárMarianTokenizer és MarianMTModel függvényeit használtuk.
- **eTranslation**: Akadémia számára ingyenesen elérhető szolgáltatás. Regisztráció után beküldtük a teszt fájlunkat, amelynek a fordítását e-mailben kaptuk meg.
- **deepL**: Az online fájl-fordító funkcióval fordítottunk, 500 mondatonként.
- **Google**: Az online dokumentum-fordító funkcióval fordítottunk, 500 mondatonként.
- **Microsoft**: Azure Translator szolgáltatás[25] felhőalapú dokumentum-fordító moduljával fordítottunk.
- **Yandex**: Az online dokumentum-fordító funkcióval fordítottunk, 500 mondatonként.

4. Eredmények

A különböző modellek, rendszerek kiértékelésére a SacreBLEU (Papineni és mtsai, 2002; Post, 2018) és a chrF (Popović, 2015) metrikákat használtuk. A BLEU metrika mellett azért választottuk a chrF metrikát, mert az karakteralapú, ami a ragozó nyelvek esetében, mint magyar nyelv, pontosabb kiértékelést eredményez. A chrF kiértékelésnél a 6-gram mellett a 3-gram értékeket is megmértük.

A 4. táblázatban láthatók a különböző gépifordítók eredményei. Az ipari alkalmazások közül az eTranslation és a deepL teljesítettek a legjobban és a közöttük lévő különbség statisztikailag nem szignifikáns mértékű. A második minőségi kategóriába a nagy cégek rendszerei (Google, Microsoft) kerültek, míg a Yandex rendszere nagymértékben elmarad ezektől.

[25] https://docs.microsoft.com/hu-hu/azure/cognitive-services/translator/document-translation/overview

forrás	- Oh, no. If you think you're tucking me away somewhere, you've got another think coming.
referencia	Ha azt tervezi, hogy bedug valahová, akkor terveljen ki valami mást.
google	- Óh ne. Ha azt hiszed, hogy elrejtesz valahova, akkor más gondolat jön.
M2M100	Ha azt hiszed, hogy valahol elrejtesz engem, van egy másik gondolkodásod.
mT5	Ha azt hiszed, hogy elrángatsz valahol, akkor jön egy másik gondolat.
BART	Ha azt hiszed, hogy el akarsz dugni valahova, akkor másra is gondolhatsz.
Marian big	Ha azt hiszed, hogy eldughatsz valahova, akkor tévedsz.

3. táblázat. Példamondat a fordító rendszerek összehasonlítására.

	BLEU	chrF-3	chrF-6	Tanítási idő
Marian big	**37,30**	61,61	56,80	21 nap
BART	**36,89**	60,77	56,48	43,5 nap
mT5	27,69	53,73	48,57	26 nap
Marian small	26,99	51,31	46,07	9 óra
Helsinki en-hu	27,21	55,03	49,82	-
Helsinki en-fiu	24,23	52,68	47,16	-
Helsinki en-urj	24,16	52,56	47,09	-
Helsinki en-multi	14,39	43,69	36,74	-
M2M100 - 1.2B	21,62	50,93	45,73	-
M2M100 - 418M	18,75	48,40	42,72	-
eTranslation	28,29	56,00	51,27	-
deepL	26,54	56,06	51,01	-
Google	25,30	54,09	49,06	-
Microsoft	25,22	53,02	48,00	-
Yandex	19,22	49,78	43,94	-

4. táblázat. A gépifordító rendszerek eredményei.

A kutatásra szánt modellek közül egyértelműen a Helsinki en-hu modell teljesített a legjobban, ami nem meglepő, hiszen a tanítóanyag szempontjából átfedés volt a használt tanítóanyaggal, továbbá kétnyelvű és nem több nyelvű modell. Az M2M100 nagyobbik modellje, annak ellenére, hogy 100 nyelvet tud, versenyképes eredményt ért el angol-magyar nyelvpárra.

Az általunk tanított modellek közül a Marian big és a BART modellek kerültek a legjobb minőségi osztályba szinte azonos eredménnyel. Az elvártaknak megfelelően a Marian big modellje ért el legjobb eredményt köszönhetően annak, hogy a legtöbb paraméterrel rendelkező hálózattal dolgozik. Másfelől az implementációjának köszönhetően gyorsabban érte el ezt az eredményt, mint vetélytársai. Mindezek ellenére érdemes megemlíteni a BART modellt is mivel egy base modellként sikerült összemérhető teljesítményt nyújtania a sokkal nagyobb hálózattal rendelkező társával szemben.

Az mT5 modellünk erőforrás hiányában csak 1 epochon finomhangoltuk, ezért nem sikerült olyan magas eredményt elérni vele. Azonban így is versenyképes teljesítményt nyújtott, megelőzve a legtöbb ipari és kutatási modellt, annak el-

lenére, hogy az mT5 vegyes feladatokra tanították elő és csak 1-szer látta a teljes anyagunkat.

A 3. és az 5. táblázatokban 1-1 példamondat olvasható, ahol az érdekesebb rendszerek fordításait emeltük ki. A fordításokat vizsgálva látható, hogy mindegyik rendszer kimenetén viszonylag olvasható szövegek voltak, továbbá a köztük lévő eltérések nagyrészt nyelvtani szerkezetbeli különbségek. Ez a jelenség a példamondatokban is megmutatkozik: a hibás fordítások fő oka a ragozásból fakadó tartalmi eltérés. A bemutatott példákon megfigyelhető, hogy a BART és a Marian big modellek fordításai adják vissza legjobban a forrásmondatok mondanivalóját annak ellenére, hogy ez nem felel meg karakter szinten a referencia mondatnak.

forrás	This may not make much sense to you, sir, but I'd like to ask your permission to date your daughter.
referencia	Szeretném megragadni az alkalmat uram, hogy az engedélyét kérjem, hogy találkozhassak a lányával.
google	Lehet, hogy ennek nem sok értelme van, uram, de szeretném engedélyét kérni a lányával való randevúzáshoz.
M2M100	Lehet, hogy ez nem sok értelme, uram, de szeretném kérni az engedélyét, hogy dátumot a lánya.
mT5	Talán nem sok értelme van, uram, de szeretném kérni az engedélyét, hogy randizzon a lányával.
BART	Lehet, hogy önnek nincs sok értelme, uram, de szeretném az engedélyét kérni, hogy randizhassak a lányával.
Marian big	Ennek talán nincs sok értelme, uram, de szeretném az engedélyét kérni, hogy randizhassak a lányával.

5. táblázat. 2. példamondat a fordító rendszerek összehasonlítására.

5. Összegzés

Kutatásunkban különböző neurális gépifordító modelleket és rendszereket tanítottunk be és próbáltunk ki angol-magyar nyelvpárra. Egyaránt kísérleteztünk kutatásban használt és iparban alkalmazott gépifordító módszerekkel, rendszerekkel is. Kísérleteink során a meglévő modellek mellett saját gépifordító rendszereket is tanítottunk. Betanítottunk kettő Marian NMT rendszert, egy kicsi és egy nagy modellt. Továbbá betanítottunk egy saját BART modellt, amelyet ezután finomhangoltuk gépi fordításra. Végül egy előtanított mt5 modellt finomhangoltuk angol-magyar gépi fordításra. Eredményeinkben megmutattuk, hogy az általunk tanított nagy Marian NMT modell és a BART modell szignifikánsan magasabb eredményt értek el az összes többi modellhez képest. Kettőjük versenyében a BART minimális értékkel marad csak le a Marian Big modelltől, ami érdekes eredmény, mivel a BART kevesebb paraméterekkel volt képes versenyképes eredményt elérni.

Köszönetnyilvánítás

A publikációban szereplő kutatást, amelyet a Pázmány Péter Katolikus Egyetem és az IdomSoft Zrt. valósított meg, az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta az Infokommunikációs és Információtechnológiai Nemzeti Laboratórium keretében.

Hivatkozások

Aharoni, R., Johnson, M., Firat, O.: Massively multilingual neural machine translation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3874–3884. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics 7, 597–610 (Mar 2019)

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., MĂźller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019)

Bucila, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 535–541. KDD '06, Association for Computing Machinery, New York, NY, USA (2006)

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://aclanthology.org/D14-1179

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: ICML. pp. 4411–4421 (2020)

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (July 2018)

Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (Oct 2014)

Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020)

Li, L., Jiang, X., Liu, Q.: Pretrained language models for document-level neural machine translation (2019)

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics 8, 726–742 (11 2020)

Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. In: 5th International Conference on Learning Representations. Palais des Congrès Neptune, Toulon, France (2017)

Miculicich, L., Ram, D., Pappas, N., Henderson, J.: Document-level neural machine translation with hierarchical attention networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020)

Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002), https://aclanthology.org/P02-1040

Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp.

392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015), https://aclanthology.org/W15-3049

Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Brussels, Belgium (Oct 2018)

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: Unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 6639–6649. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)

Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140), 1–67 (2020)

Rescigno, A.A., Monti, J., Way, A., Vanmassenhove, E.: A case study of natural gender phenomena in translation: A comparison of Google Translate, bing Microsoft translator and DeepL for English to Italian, French and Spanish. In: Workshop on the Impact of Machine Translation (iMpacT 2020). pp. 62–90. Association for Machine Translation in the Americas, Virtual (Oct 2020)

Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (Aug 2016)

Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., Liu, T.Y.: Multilingual neural machine translation with knowledge distillation. In: Proceedings of the 7th International Conference on Learning Representations (ICLR 2019). New Orleans, LA, USA (2019)

Tiedemann, J., Thottingal, S.: OPUS-MT — Building open translation services for the World. In: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT). Lisbon, Portugal (2020)

Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)

Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: CCNet: Extracting high quality monolingual datasets from web crawl data. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (May 2020)

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine

translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144 (2016)

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498. Association for Computational Linguistics, Online (Jun 2021), https://aclanthology.org/2021.naacl-main.41