Soknyelvpáros gépi fordítás hatékony és megbízható kiértékelése

Oravecz Csaba, Sass Bálint, Tihanyi László

MTA Nyelvtudományi Intézet
e-mail:{oravecz.csaba,sass.balint,tihanyi.laszlo}@nytud.hu

Kivonat Gépi fordítások kiértékelésére a legmegbízhatóbb módszer az emberi szakértői kiértékelés, mely egyértelműen elsődleges mindenfajta egyéb megközelítéssel szemben. A dolgozat arra keresi a választ, hogy milyen elfogadható alternatívákkal váltható ki a szakértői kiértékelés abban az esetben, amikor ez a preferált, ugyanakkor rendkívül erőforrásigényes módszer a kiértékelendő szövegek nagy mennyisége, illetve a kiértékelési feladat sajátos paraméterei miatt nem alkalmazható. A javasolt megoldás a rendelkezésre álló többféle típusú kiértékelési információt rugalmasan kombináló és ennek alapján minőségi klasztereket képző eljárás, ahol az egyes klasztereken belül minden fordítási kimenethez véletlenszerűen generálódik az aktuális rangsor.

Kulcsszavak: gépi fordítás, fordításkiértékelés, korreláció, fordítóportál

1. Bevezetés

A kutatás hátterét az iTranslate4.eu nemzetközi projektum adja, melynek keretében elkészült egy 63 nyelvpár közötti automatikus gépi fordítást és egyéb fordításon alapuló szolgáltatást kínáló webportál. A weboldalon a fordítást 14 szolgáltató által kifejlesztett szabályalapú, illetve statisztikus fordítómotorok végzik. A 63 nyelvpár összesen $63 \times 62 = 3906$ nyelvpár közötti fordítást tenne szükségessé. Bár a portál számára valójában csak 233 nyelvi motor áll rendelkezésre, megfelelő közvetítő nyelvek megválasztásával a portál kiszolgálja valamennyi nyelvi irányt, így tetszőleges nyelvről tetszőleges másikra fordít.

A portál egyedi sajátossága hasonló online fordítókkal szemben, hogy egy-egy kérésre több megoldással is tud szolgálni. Mind a különböző programok gyártóinak, mind a felhasználóknak természetes igénye, hogy ezek az alternatívák minőségi sorrendben jelenjenek meg. Ehhez szükség van az egyes fordítók kérdéses nyelvpárok szerinti teljesítményének a kiértékelésére, hatékony és megismételhető, a fordítómotorok minőségi változását követni képes módon. A feladat volumenének következtében a szakértői emberi kiértékelés nem vehető számításba, más módszereket kell kidolgozni. A kiértékelési feladat célja tehát alapvetően bekezdés hosszúságú szövegek sorrendbe rendezése, amelynél figyelembe kell venni, hogy

– a minősítés nem lassíthatja a fordítási folyamatot,
– a szövegek megjelenítésének célja a megértés és nem az újrafelhasználás, ezért olyan offline kiértékelési eljárások preferálandók, amelyek inkább a felhasználói vélemény, mintsem az esetleges utószerkesztéshez szükséges költségmetrika alapján rangsorolnak.

Az offline megoldással természetesen nem az éppen megjelenő fordításokat rangsoroljuk, hanem az azokat létrehozó fordítóprogramokat. A rangsor a fordítóprogramok szempontjából releváns, hiszen a következő kiértékelésig meghatározza azok sorrendjét. A minősítések a fordításokkal együtt nem jeleníthetők meg, hiszen a felhasználó a konkrét megoldás minősítését várná el, a fordítók általános minősítése ezt pedig csak közelítheti.

2. Gépi fordítások kiértékelése

A gépi fordítások kiértékelése közismerten körülményes és bonyolult feladat, melyre hosszú ideje keresnek hatékony és könnyen kivitelezhető megoldást. Az automatikus kiértékelő metrikák legismertebbje, a Bleu-mérték [17] mellett mára további módszerek sokaságát fejlesztettek ki (lásd pl. a [7] kiadványt, illetve a [4] tanulmányban található összefoglalót). Széles körben elfogadott ugyanakkor, hogy az automatikus módszerek megbízhatósága jelentősen elmarad a (szakértői) humán kiértékeléstől [4], ezért gyakorlati hasznuk leginkább a fordítómotorok fejlesztése során van [6]. A legjobb eredményeket adó eljárások ezen túl olyan nyelvi előkészítést és adott nyelvi erőforrások (pl. WordNet) meglétét igénylik, melyek a jelen feladat kontextusában nyilvánvalóan a kérdéses nyelvek nagy részében nem állnak rendelkezésre. További probléma, hogy a statisztikai alapú fordítórendszerek, melyek egyre inkább dominánsak a szabályalapú rendszerek felett, egyre több, gyakorlatilag minden elérhető adatot igyekeznek felhasználni betanításuk érdekében. Ezért lehetetlen, de legalábbis bizonytalan kimenetelű egy elfogulatlan, fenntartható és folyamatos nagy léptékű kiértékelő környezetet kifejleszteni, hiszen a tesztadatok függetlensége nem biztosítható.

A fentiek fényében egyértelmű a humán kiértékelés elsődlegessége akkor, amikor a feladat a többféle fordítómotor által szolgáltatott fordítások valamilyen rangsorba állítása. A legjobb megoldás természetesen a szakértői kiértékelés, ám az így kapott eredmények objektív értelmezése sem problémamentes [2]. Kézenfekvő persze, hogy jelen esetben ez a rendkívül erőforrásigényes módszer a kiértékelendő szövegek nagy mennyisége, illetve a kiértékelési feladat sajátos paraméterei miatt eleve szóba sem jön, a végső megoldásban fenntartható módon nem alkalmazható.

3. Módszerek és vizsgálatok

3.1. A kiértékelendő nyelvek, nyelvpárok és fordítómotorok

Bár 63 nyelv esetén a nyelvpárok elvi kombinációjának száma 3096, ennél jóval kevesebb nyelvpár kiértékelésével kellett foglalkoznunk. Ennek több oka is volt:

egyrészt a valójában nyelvi motorral is támogatott nyelvpárok száma csak 233, a többi esetben pedig közvetítő nyelven keresztül két lépésben fordít a rendszer. A portálunkhoz hasonlóan a Google és a Microsoft fordítóprogramjai is közvetítő nyelvet használnak, azaz az általuk támogatott nyelvpárok száma ezek esetén is csak a nyelveik számának a kétszerese. A többi 12 fordítóprogram a minőségi normák betartása érdekében nem végez közvetítő nyelves fordítást, itt a nyelvpárok száma közvetlenül ismert. Mivel a kiértékelési feladatunk célja rangsorolás volt, ezért nem kellett figyelembe venni azokat a nyelvpárokat sem, amelyeken csak egy versenyző indult, ezzel a nyelvpárok száma 106-ra csökkent.

A weboldalon fordító programok két nagy kategóriába csoportosíthatók. Az egyikbe a szerződéses partnerek, a másikba pedig a Google és a Microsoft tartoznak. Az utóbbiak szabadon elérhető programozói felület (API) segítségével integrálhatók. Mivel azonban mind a Google, mind a Microsoft fordítók ilyen jellegű felhasználása hamarosan fizetős szolgáltatássá válik, ezért ezeknek a nyelvpároknak üzemeltetése és kiértékelése csupán tájékoztató jellegű eredménnyel szolgálhat, a végleges megoldásban nem játszik szerepet. A 12 partnerfordítóból a legalább kettő által támogatott nyelvpárok száma 58 volt. Mivel a kiértékelési eljárások költségét alapvetően a kiértékeléshez szükséges nyelvi erőforrások (párhuzamos szövegek gyűjtése, tesztek összeállítása) teszik ki, ezek csak egy-egy új nyelvpár esetén jelentenek többletköltséget. Vagyis a partnerek miatt kiértékelendő nyelvpárok esetén a kiértékelés további költség nélkül kiterjeszthető a Google és Microsoft fordítókra is.

A kiértékelési feladat során a versenyzők számának alakulása és a különböző nyelvpárok (nyelvek ISO kód szerinti rövidítésével) az alábbiak voltak:

8: fr-de, en-de, de-fr, de-en
7: fr-en, en-fr
6: it-en, es-en, en-it, en-es
5: ru-en, pt-en, pl-en, fr-es, es-fr, es-de, en-ru, en-pt, en-pl, de-es
4: zh-en, uk-en, tr-en, sv-en, sl-en, ru-pl, ru-fr, ru-de, pl-ru, pl-fr, pl-de, no-en, lv-en, it-fr, it-es, it-de, hu-en, fr-ru, fr-it, fi-en, es-it, en-zh, en-tr, en-sv, en-lv, en-hu, en-fi, en-da, de-ru, de-pl, de-it, da-en, bg-en

A fenntartható kiértékeléshez kétféle kivitelezhető megközelítés választható, ám mindegyik felvet számos olyan kérdést, melyet a hatékony módszer kidolgozása érdekében meg kell válaszolni:

A. Valamilyen sztenderd mérték(ek) szerinti automatikus, gépi kiértékelés.
B. Emberi, de nem szakértői kiértékelés, amely nagy léptékben is alkalmazható.

3.2. Automatikus kiértékelés

Az automatikus kiértékelés (a továbbiakban AU) során az IQMT [12] keretrendszer által szolgáltatott 5 féle sztenderd mérték normalizált átlagát használtuk: BLEU [17], NIST [9], GTM [16], METEOR [1] és ROUGE [13]. Ideális esetben 3 humán referenciafordítás szükséges a kiértékeléshez, tekintve azonban a projektben szereplő nyelvek széles skáláját, ilyen mennyiségű fordítás beszerzése,

előállítása reménytelen, így egy referenciafordítást alkalmaztunk, és a felhasznált szövegek műfajának és forrásának variábilitásával próbáltuk kiegyensúlyozottabbá tenni az automatikus kiértékelést. A kívánt nyelvi erőforrások az EU párhuzamos hírkorpuszból származnak, 13 különböző témakategóriából, mintegy 80 ezer szövegszó méretben. Természetesen, hiába saját gyűjtésről van szó, itt is felmerül a források függetlenségének kérdése: vajon ezek a szövegek nem alkották-e a részét a statisztikus fordítóprogramok tanítókorpuszának.

3.3. Emberi, nem szakértői kiértékelés a Mechanical Turk rendszerben

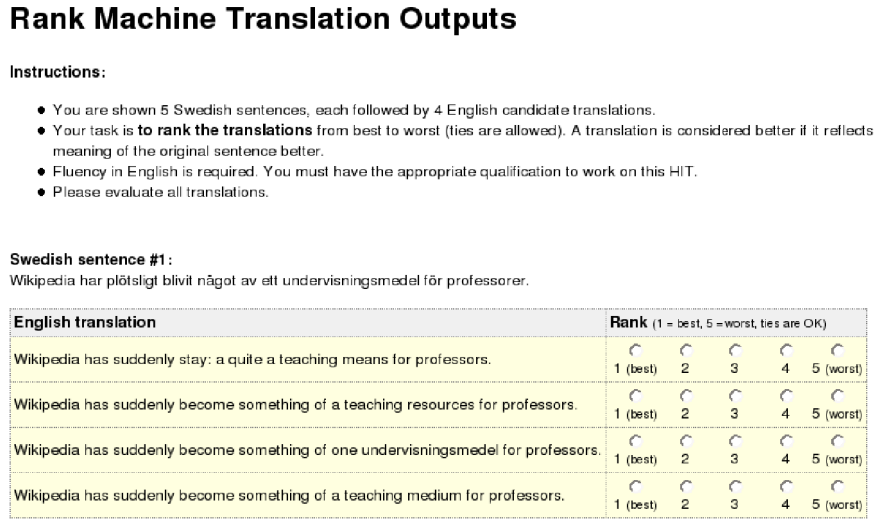
A nagyobb volumenű emberi, nem szakértői fordításértékelés megvalósítására lehetőséget adnak az utóbbi években létrejött, online elérhető *crowdsourcing* rendszerek. Ezekben a rendszerekben internetes űrlap formájában megfogalmazható, emberi intelligenciát igénylő feladatok (HIT, human intelligence task) tehetők közzé. A feladatokat a regisztrált dolgozók (worker) meghatározott fizetség ellenében végzik el. Lehetőség van a dolgozók előzetes szűrésére, például megtehetjük, hogy csak olyan dolgozók jelentkezését fogadjuk, akik már korábban adott számú HIT-et sikeresen megoldottak. A nem megfelelő minőségűnek ítélt munkavégzés esetén a fizetség visszatartható. Ezek az eszközök segítenek a munkavégzés általános minőségi szintjét magasan tartani. A *crowdsourcing* rendszerekkel tehát olcsón és gyorsan lehet megbízható minőségű megoldást találni emberi intelligenciát igénylő feladatokra [3], ugyanakkor legújabban már az ilyen rendszerek esetleges kockázataira is felhívják a figyelmet [11].

Eljárásunk. A gépi fordítások emberi, nem szakértői kiértékelésére (a továbbiakban MT) a Mechanical Turk (`http://www.mturk.com`) internetes rendszert alkalmaztuk.

Forrásnyelvenként 30 darab, téma szerint minél változatosabb közepes hosszúságú (legnagyobbrészt 10–30 szavas) mondatot gyűjtöttünk. Ezeket a mondatokat a rendelkezésre álló fordítóprogramok mindegyikével lefordítottuk. Hogy egy kiértékelési feladat ne legyen túl időigényes, egy feladatba (HIT-be) 5 mondatot tettünk, azaz a 30 mondatot 6 db 5-ös csoportra osztottuk. Egy kiértékelőnek tehát egy feladat keretében 5 db mondat fordításait kellett értékelnie.

A kiértékelőknek az a feladata, hogy 1-től 5-ig terjedő skálán minőség szerint *pontozzák* a fordításokat. Az instrukciók és egy mintafeladat – svéd–angol nyelvpárra, ahol 4 különböző automatikus fordító van – a 1. ábrán látható. A feladat a fordítások sorba rendezése, 1-től (legjobb) 5-ig (legrosszabb) skálán adott pontszám segítségével. Több mondatnak adható azonos pontszám, és a fordítások számától függetlenül 1-től 5-ig terjedő skálát használunk.

A rendszer működéséből adódóan egy kiértékelő tetszőleges számú mondat kiértékelését elvégezhette (azaz akár az összes 30 mondatét is). Ezért – hogy semmiképp se csak egy dolgozó véleményére támaszkodjunk – minden mondatot 3 különböző kiértékelővel értékeltettünk ki. Itt a különbözőséget szintén a rendszer biztosítja. Végeredményben tehát fordítónként $3 \times 30 = 90$ kiértékelési pontszámot kaptunk, ami minimum három különböző kiértékelőtől származott.



Rank Machine Translation Outputs

Instructions:

- You are shown 5 Swedish sentences, each followed by 4 English candidate translations.
- Your task is **to rank the translations** from best to worst (ties are allowed). A translation is considered better if it reflects the meaning of the original sentence better.
- Fluency in English is required. You must have the appropriate qualification to work on this HIT.
- Please evaluate all translations.

Swedish sentence #1:
Wikipedia har plötsligt blivit något av ett undervisningsmedel för professorer.

English translation	Rank (1 = best, 5 = worst, ties are OK)
Wikipedia has suddenly stay: a quite a teaching means for professors.	1 (best) 2 3 4 5 (worst)
Wikipedia has suddenly become something of a teaching resources for professors.	1 (best) 2 3 4 5 (worst)
Wikipedia has suddenly become something of one undervisningsmedel for professors.	1 (best) 2 3 4 5 (worst)
Wikipedia has suddenly become something of a teaching medium for professors.	1 (best) 2 3 4 5 (worst)

1. ábra. Egy a Mechanical Turk rendszerben megvalósított fordításkiértékelési feladat dolgozóknak szóló felülete a svéd–angol nyelvpár esetén.

A kapott 90 db érték összesítésére kétféle mérőszámot alkalmaztunk. Egyrészt egyszerűen átlagot számoltunk, másrészt az EuroMatrix projektben [5, 3.1 rész] alkalmazott mértéket használtuk, miszerint egy fordítórendszer minden olyan esetben kap egy pontot, ha egy kiértékelő szerint egy másik rendszernél jobb (vagy vele egyforma), és végül pontszám szerint rendeztük a fordítórendszereket. A két mérőszám lényegében minden esetben ugyanazt az értéket adta, ezért a pontszámok átlagával dolgoztunk a továbbiakban.

Minőségbiztosítás. A fordításértékelési feladat megoldásához nyilván szükséges mindkét nyelv megfelelő ismerete, magasszintű ismeret főként a célnyelv esetében kívánatos. Annak érdekében, hogy valóban jó minőségű értékeléseket kapjunk, bevezettük azt, hogy a dolgozóknak először ki kell tölteniük egy rövid tesztet az adott nyelvpárra vonatkozóan, és csak akkor dolgozhatnak a kiértékelésben, ha ez jó eredményű. A Mechanical Turk terminológiájával egy megfelelő minősítés (qualification) meglétét követeljük meg, mielőtt a dolgozó hozzákezd a munkához.

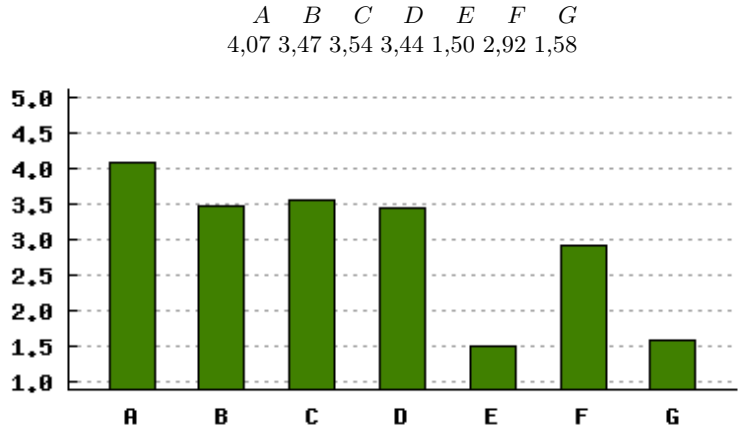
A célnyelvre fordítás képességét egy négy kérdésből álló teszttel mértük, négy darab forrásnyelvi mondat esetében kellett megmondani, hogy a felkínált fordítások közül melyik a legjobb. A szándékosan hibás fordításokban morfológiai, szintaktikai és szemantikai, szókincsbeli hibák egyaránt előfordultak.

Sorrendkeverés. Kutatásunk első szakaszában a fordítások mindig fix sorrendben jelentek meg. Ez a sorrendből adódó nem kívánt torzító hatáshoz vezetett.

E hatást és kiküszöbölését a német–angol nyelvpáron mutatjuk be, ahol 7 fordítórendszert teszteltünk.

A pszichológiában ismert az a jelenség, hogy ha több azonos típusú entitást kell értékelnünk, akkor jelentősége van annak, hogy ezek a bizonyos értékelendő dolgok milyen sorrendben kerülnek elénk. Megfigyelték, hogy bizonyos esetekben hajlamosak vagyunk az elsőként látottat előnyben részesíteni (*primáciahatás*, vö. [15]), más feltételek mellett pedig az utolsót (*recenciahatás*, vö. [8]). Ezek a jelenségek főként akkor figyelhetők meg, mikor az adott jelölt megfigyelése után azonnal értékelni kell, nem várhatjuk meg a pontszámokkal az összes versenyzőt (ilyen például a műkorcsolya-zsűrizés struktúrája). Esetünkben lehetőség volt a jelöltek (fordítások) többszöri vizsgálatára, összevetésére, és csak az összes jelölt vizsgálata után kellett döntést hozni, mégis határozott primáciahatást találtunk, amit torzította az eredményeket.

A német–angol nyelvpáron végzett első kísérletekben tehát a 7 angol fordítás mindig fix sorrendben, a fordítórendszerek neve szerinti betűrendben jelent meg az eredeti német mondat után. A fordítónként 90 értékből adódó átlagos pontszámok a 2. ábrán láthatók.

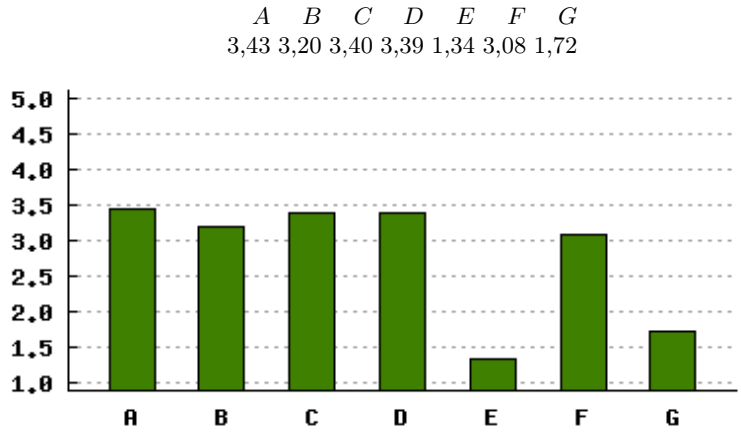


2. ábra. Fordítókénti átlagos pontszámok. Itt a 7 angol fordítás mindig *a fordítók neve szerinti betűrendben* következett az eredeti német mondat után. (Az osztályzás itt eredetileg 1-től 7-ig történt, utólag normáltuk ezt az összehasonlíthatóság kedvéért az 1..5 skálára a következő módon: normált = eredeti $\times \frac{2}{3} + \frac{1}{3}$.)

A sorrendi hatások kiegyenlítése nem mindig könnyű [8], esetünkben azonban egy egyszerű, determinisztikus *sorrendkeverő* algoritmus segítségével biztosítani lehetett azt, hogy minden pozíció esetében igaz legyen az a feltétel, hogy minden fordító ugyanannyiszor fordul elő az adott helyen.

A sorrendkeverő algoritmus alkalmazásával a fordítások determinisztikus módon változó, a keverőalgoritmus által meghatározott sorrendben követték egy-

mást. A német–angol nyelvpár esetében a fordítókénti 90 értékből így adódó átlagos pontszámokat a 3. ábrán láthatjuk.



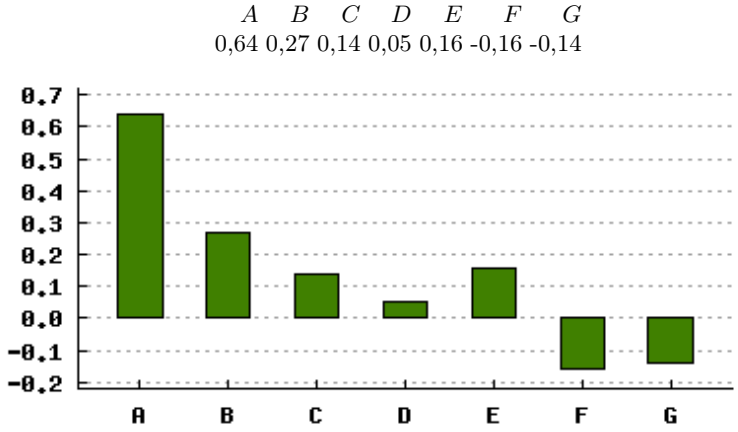
3. ábra. Fordítókénti átlagos pontszámok. Itt a 7 angol fordítás mindig *változó, a keverőalgoritmus által meghatározott sorrendben* következett az eredeti német mondat után.

A 2. és a 3. ábrát összevetve látjuk, hogy egy helyen maga a sorrend is megváltozott (*B*-*D* helyett *D*-*B*), de ennél lényegesebb annak feltárása, hogy milyen mértékben változtak a pontszámok a két elrendezés között. A különbségeket ábrázoltuk a 4. ábrán. Az ábra tanúsága szerint egyértelmű primáciahatást tapasztalunk („a fix első hely jogtalan előnnyel jár; aki előrébb van, az érdemtelenül több pontot kap”), egyfajta fordított recenciahatással erősítve („aki hátrébb van, az igazságtalanul kevesebb pontot kap”). A torzító hatás arányos az eredeti pozícióval.

Az eredmény arra hívja fel a figyelmet, hogy az ilyenfajta többszöri értékeléses feladatokban egyáltalán nem mindegy, hogy milyen sorrendben szerepelnek az értékelendő entitások, a sorrend nagyban befolyásolja az eredményt. Az igazságos értékeléshez fontos a sorrendi hatások kiküszöbölése, különben torzul az eredmény.

3.4. Felhasználói visszajelzések

A harmadik kiértékelő komponenst a felhasználói visszajelzések (továbbiakban FV) alkotják. Ezek valójában az egyes fordításokra érkezett szavazatok, amelyeket a portálon adhatnak le a felhasználók. Egy fordítás esetén több megoldás is megjelölhető. A szavazatokat a portál megnyitása óta gyűjtjük. Bár a szavazati hajlandóság viszonylag magas (5%-os), az induló weboldal látogatóinak alacsony száma miatt az adatok mennyisége csak lassan nő. A szavazás során



4. ábra. Fordítókénti átlagos pontszámok *különbsége* az első – sorrendi hatásnak kitett (vö. 2. ábra) –, és a második – sorrendi hatásra semleges (vö. 3. ábra) – elrendezés között. Bár az eltérés csak A esetében szignifikáns (kétmintás Welch-próba: $p \ll 0.05$), jól látható egy trend, miszerint a sorrendi hatásnak kitett esetben az előrébb lévők jogtalan előnyhöz jutnak, a hátrébb lévők pedig hátrányt szenvednek.

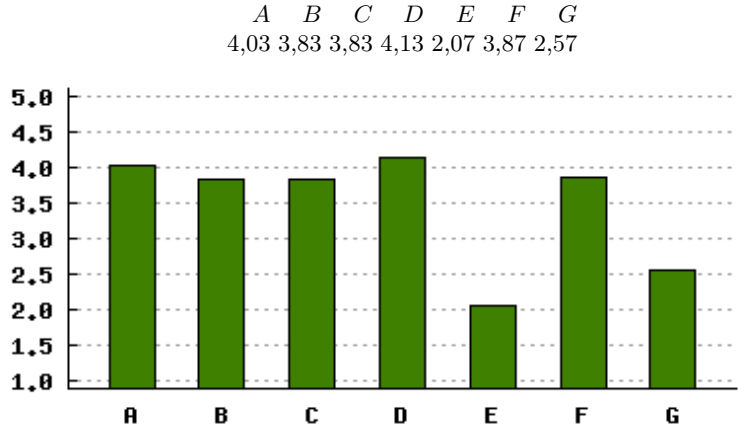
kiderült, hogy a közvetítőnyelves megoldások is használatban vannak, és szavazatokat tudnak gyűjteni. Ezekre sem az automatikus, sem az MT kiértékelések erőforrás hiányában nem tudtak adatokkal szolgálni. A partnerek között elvi egyetértés alakult ki arról, hogy a jövőben, megfelelő mennyiségű adat birtokában az FV kiértékelés legyen elsődleges, hiszen ez elvileg valódi fordítási kérdésekre valódi felhasználók által adott értékelést képvisel. Vizsgálni kell azonban ennek a kiértékelési módszernek a megbízhatóságát is.

4. Eredmények

4.1. A kiértékelések összevetése

Fontos kérdés, hogy a 3.3. részben leírt módszer segítségével a Mechanical Turk rendszerrel valóban lehetséges-e magas megbízhatóságú kiértékelést végezni. Ezt úgy vizsgálhatjuk meg, hogy a szakértő véleményét vetjük össze a nem szakértő dolgozók véleményével. Ennek érdekében kiértékeltettük a már említett német–angol nyelvpárt egy szakértővel. A szakértő által adott 30 darab pontszám átlagos értéke a 5. ábrán látható.

Annak ellenére, hogy a kis eltérések miatt a fordítók sorrendjében lényeges különbségek vannak, megfigyelhető, hogy a nem szakértői kiértékelők (3) és a szakértő (5) meglehetősen hasonlóan értékelte a fordításokat, ahogy a két ábrán látható grafikon lefutásán is látható. Célszerű ezért a rangsorok összehasonlítására szokásosan használt Spearman-féle rangkorrelációs együttható helyett más



5. ábra. A szakértő átlagos pontszámai német–angol nyelvpárra. A grafikon lefutása lényegében megegyezik a 3. ábrán láthatóval.

megközelítést alkalmazni a hasonlóság mértékére. Kolmogorov–Szmirnov próbával vizsgáltuk meg, hogy mennyire valószínű, hogy a két grafikon ugyanazt írja le. A p értékre 0,05-nek adódott, azaz 5% hiba mellett mondhatjuk, hogy igaz az, hogy a nem szakértők és a szakértő gyakorlatilag ugyanúgy értékelte a fordításokat. Emiatt a Mechanical Turk rendszerben kapott kiértékeléseket is megbízhatónak tarthatjuk, azaz általánosságban támaszkodhatunk erre a sokkal olcsóbb és egyszerűbben kivitelezhető emberi kiértékelési metódusra. Korábban úgy gondolták [3], hogy a *crowdsourcing* megbízható kiértékelési eredményeket ad, ez később megkérdőjeleződött [4], jelen eredményeink azt mutatják, hogy ha az alkalmas dolgozókat a 3.3. részben bemutatott eljárás segítségével választjuk ki, a megbízhatóság megfelelő szintű lesz.

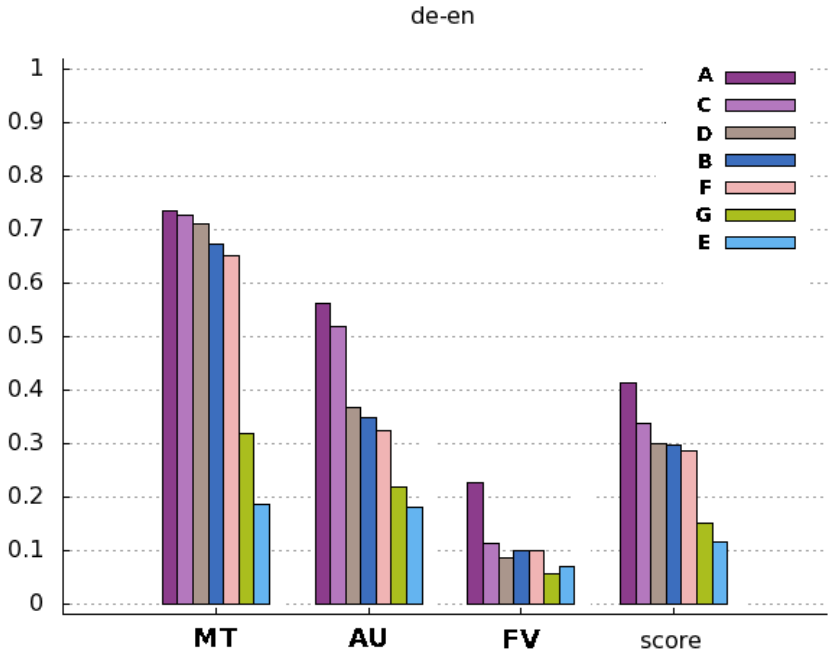
A további komponensek összehasonlítása során beigazolódott, hogy a szakértői kiértékeléshez legközelebb álló MT módszer után a felhasználói visszajelzések a legmegbízhatóbbak, az automatikus kiértékelés pedig, különösen a statisztikai fordítókkal szembeni elfogultság miatt a legkevésbé megbízható. Azokon a nyelvpárokon, ahol közvetett és közvetlen fordítások is elérhetők voltak, egyértelműen megmutatkozott az utóbbiak minőségi fölénye.

4.2. Javasolt kiértékelési módszer

A gyakorlati alkalmazásban nehezen védhető egy, a kiértékelések alapján rögzített rangsorba rendezés a fordítómotorok között, és a fordítások e szerinti megjelenítése. A 6. ábra illusztrál egy olyan összevont rangsort, ahol az egyes fordítómotorokhoz rendelt mérték (*score*) a három komponens (s) súlyozott átlaga ($w_1 = 0.1, w_2 = 0.3, w_3 = 0.6$):

$$score = \frac{w_1 s_{\mathrm{AU}} + w_2 s_{\mathrm{MT}} + w_3 s_{\mathrm{FV}}}{3} \tag{1}$$

A kis minőségi különbséggel hátrább sorolt partner jogosan tiltakozik, hogy a



6. ábra. Az egyes komponensek eredményei és az összevont rangsor.

sohasem 100%-osan megbízható értékelés(ek) alapján *véglegesen* rosszabb helyre kerül. Ezért a rögzített rangsor helyett az alábbi javasolt módszerrel próbáljuk kiküszöbölni ezt a problémát.

Képezzünk az értékelés során kapott eredmények alapján a fordítómotorok között minőségi klasztereket. A klaszterek számát az értékeléskor kapott adatok alapján kell automatikusan meghatározni (a 3., 5. és 6. ábrán látható adatok alapján például két minőségi klasztert célszerű képezni, ha eltekintünk az AU módszer elfogultságától a statisztikus fordítók felé). Erre kétféle megközelítés alkalmazható: a klaszterek számát előre megkívánó algoritmus (pl. k-means) esetében valamilyen segédalgoritmus (lásd pl. [14,18]), illetve a klaszterek számát is meghatározó klaszterező algoritmus [10]. Az egyes klasztereken belül alapesetben véletlen rendezés szerint jelennek meg a fordítások. A klaszterek képzéséhez szükséges bemenő adatot az adott nyelvpárra kétféleképpen állíthatjuk elő. Egyrészt a rendelkezésre álló kiértékelő komponensek eredményeinek például (1) szerinti összevonásával, vagy az éppen legmegbízhatóbbnak tekinthető és elegendő adatot szolgáltató komponens kizárólagos figyelembevételével (ahol a megbízhatósági sorrend a következő MT $\rightarrow$ FV $\rightarrow$ AU). A legjobb megoldás kiválasztásához

további értékelési adatok és vizsgálatok szükségesek, ahol természetesen azt is meg kell határozni, mit fogadunk el elegendő adatnak.

Ez a módszer feltétlen igazságosabb és a partnerek által is elfogadhatóbb, mint a kötött rangsor alapján történő rendezés, megvalósítása azonban technikai okok miatt csak részleges lehet. A fordítómotorok eltérő sebessége miatt portál felületen definiált meghatározott maximális válaszidő (jelenleg 1mp) már eleve kialakít egy sorrendet. A portál szolgáltatásait közvetítő API alkalmazásokban pedig a hívó fél állítja be a kért megoldásokat, az általa tapasztalt sebességi és minőségi eredmények alapján.

5. Összefoglalás és további feladatok

A tanulmányban megvizsgáltuk, hogy egy konkrét alkalmazásban hogyan valósítható meg gépi fordítások kiértékelése olyan környezetben, ahol számos gyakorlati paramétert kell figyelembe venni. Javaslatot tettünk olyan kiértékelési módszerre, amely választ ad a felmerülő problémákra: megbízható, fenntartható és soknyelvpáras fordítás esetén is alkalmazható, ezzel együtt védhető és igazságos minősítést eredményez. A portál működése során gyűjtött adatok mennyiségének növekedése további részletes vizsgálatok elvégzésére ad lehetőséget, melyek kiértékelése még megalapozottabban kimutathatja az egyes fordítók közötti minőségi különbségeket.

Hivatkozások

1. Banerjee, Satanjeev és Lavie, Alon. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005, 65–72.
2. Bojar, Ondřej, Ercegovčević, Miloš, Popel, Martin és Zaidan, Omar. A Grain of Salt for the WMT Manual Evaluation. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland. Association for Computational Linguistics, July, 2011, 1–11.
3. Callison-Burch, Chris. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, 2009, 286–295.
4. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Peterson, Kay, Przybocki, Mark és Zaidan, Omar. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden. Association for Computational Linguistics, July, 2010, 17–53.
5. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof és Schroeder, Josh. Findings of the 2009 Workshop on Statistical Machine Translation. In: *Proceedings of the EACL Workshop on Statistical Machine Translation*, 2009, 1–28.
6. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof és Zaidan, Omar. Findings of the 2011 Workshop on Statistical Machine Translation. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland. Association for Computational Linguistics, July, 2011, 22–64.

7. Callison-Burch, Chris, Koehn, Philipp, Monz, Christof és Zaidan, Omar F. szerk. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, July, 2011.
8. de Bruin, Wändi Bruine. Save the Last Dance for Me: Unwanted Serial Position Effects in Jury Evaluations. *Acta Psychologica*, 2005, 118:245–260.
9. Doddington, George. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *HLT-01*, 2002.
10. Ester, Martin, Peter Kriegel, Hans, S, Jörg és Xu, Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, 1996, 226–231.
11. Fort, Karën, Adda, Gilles és Cohen, K. Bretonnel. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 2011, 37(2):413–420.
12. Giménez, Jésus. *IQMT. A Framework for Automatic Machine Translation Evaluation based on Human Likeness*. TALP Research Center, 2007.
13. Lin, Chin-Yew és Och, Franz Josef. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics, 2004.
14. Lleti, R., Ortiz, M.C., Sarabia, L.A. és Sánchez, M.S. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 2004, 515(1):87 – 100. Papers presented at the 5th COLLOQUIUM CHEMIOMETRICUM MEDITERRANEUM.
15. Mantonakis, Antonia, Rodero, Pauline, Lesschaeve, Isabelle és Hastie, Reid. Order In Choice: Effects of Serial Position on Preferences. *Psychological Science*, 2009, 20(11):1309–1312.
16. Melamed, I. Dan, Green, Ryan és Turian, Joseph P. Precision and recall of machine translation. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, Stroudsburg, PA, USA. Association for Computational Linguistics, 2003, 61–63.
17. Papineni, Kishore, Roukos, Salim, Ward, Todd és Zhu, Wei-Jing. Bleu: A method for automatic evaluation of machine translation. In: *ACL-02*, Philadelphia, PA. 2002.
18. Sugar, Catherine A. és James, Gareth M. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 2003, (98):750–763.