Fordítási plágiumok keresése

Pataki Máté

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi utca 11.
pataki.mate@sztaki.hu

Kivonat: Napjainkban egyre több diák beszél idegen nyelveken, ami előny, hiszen fel tudják dolgozni az idegen nyelvű szakirodalmat és tudományos eredményeket, hátrány azonban, ha ezt hivatkozás nélkül teszik, azaz plagizálnak. Az elmúlt egy év alatt egy kutatás keretében arra kerestük a választ, hogy meg lehet-e találni, fel lehet-e ismerni a fordítási plágiumokat. Ennek során egy olyan algoritmust fejlesztettünk ki, amely képes egy nagyméretű, idegennyelvű adatbázisból kikeresni egy magyar nyelvű dokumentumban idézett, lefordított szövegrészeket.

1 Bevezetés

Természetes nyelvű szövegek fordításának megtalálása nemzetközi szinten is megoldatlan, még a sokak által beszélt angol és német nyelvek között is, ugyanakkor megoldása számos területen jelentene nagy előrelépést. A kutatási eredmények nemcsak plágiumok felkutatásában, hanem a párhuzamos korpuszok építésében, a hírek, cikkek, szövegek terjedésének a vizsgálatában, hasonló témákkal dolgozó emberek, kutatócsoportok megkeresésében is alkalmazhatók.

A párhuzamos korpuszok nagy jelentősége nemcsak az oktatásban rejlik, e korpuszok számos kutatás alapjaként, algoritmusok tanító adatbázisaként is szolgálnak. Használják őket az alkalmazott nyelvészetben: szótárkészítők, gépi fordítók számára, valamint kontrasztív nyelvészeti kutatásokhoz is elengedhetetlenek.

Európában fontos téma a plágiumkeresés, de még nemzetközi szinten is csak kutatási terület a fordítási plágiumok keresése. [1] Az irodalomban ismertetett legtöbb algoritmus nyelvpárfüggő, azaz egymáshoz nyelvtanban hasonló nyelvek esetén – barátságos nyelvpárok – jól működik, de jelentősen eltérő nyelvtanú nyelvek esetén rossz eredményt mutat. Angol-német nyelvpárra például egész szép eredményeket értek már el, míg az angol-lengyel nyelvpárra ugyanaz az algoritmus használhatatlannak bizonyult. A magyar nyelvben három fő akadály van: a) nem kötött szórend, b) ragozás, c) jelentős nyelvtani különbözőség az angol nyelvtől.

Dr. Debora Weber-Wulff kétévente teszteli az összes elérhető plágiumkeresőt, 2010-ben 48 plágiumkeresőt tesztelt, és azt állapította meg, hogy:

> „The biggest gap in all the plagiarism checkers was the inability to locate translated plagiarism.” [2]

Azaz a jelenleg elérhető plágiumkeresők egyáltalán nem foglalkoznak a fordítási plágiumok problémájával. Az első publikus eredmények többnyelvű plágiumkeresési algoritmusokról a CLEF 2010 konferencián [3] jelentek meg, de itt is csak barátságos nyelvpárokkal (angol, német, spanyol) próbálkoztak, és automatikus fordítót használtak a plágiumok megtalálására:

> „After analyzing all 17 reports, certain algorithmic patterns became apparent to which many participants followed independently. ... In order to simplify the detection of cross-language plagiarism, non-English documents in D are translated to English using machine translation (services).” [4]

2 Az algoritmus

A legtöbb szakirodalomban és kezdeti kutatásokban olyan algoritmusokat láthatunk a fordítási plágiumok keresésére, amelyek a jelenlegi egynyelvű keresések adaptálásai egy adott nyelvpárra. A legjobb plágiumkeresők átlapolódó szavas darabolást (n-gramokat) használnak a szövegek összehasonlítására, a plágiumkeresésre. [4] Ez az algoritmus szó szerinti egyezést keres, amelyet számos más algoritmussal igyekeznek javítani, hogy kisebb átírásokat, eltéréseket ne vegyen figyelembe, ezek közül a legygyakrabbak az alábbiak: a) stopszavak szűrése, b) szótövezés, c) bizonyos szavak kicserélése egy szinonimára, d) szavak sorrendezése az n-gramon belül. Ezek a változtatások sokkal nehezebbé teszik a plágiumok elrejtését, és jelentősen megnövelik a lebukás kockázatát, ugyanakkor különböző nyelven írt szövegek között még mindig nem teszik lehetővé az összehasonlítást.

Többen is próbálkoztak automatikus, gépi fordítók alkalmazásával, hogy két szöveget azonos nyelvre hozzanak, ugyanakkor ezen fordítók eredményei ma még nagyon megbízhatatlanok, nagyban függnek az adott nyelvpártól, a szöveg témájától, a mondatok összetettségétől. Összefoglalva elmondhatjuk, és ez nem csak a gépi fordítókra igaz – habár azokra kiemelten az –, hogy egy fordítás komoly változtatást eredményez a szövegen, hibákat visz be, és a szavak mondaton belüli sorrendjén is nagymértékben változtat, főleg az olyan nem kötött szórendű nyelvek esetében, mint amilyen a magyar.

A gépi fordítókat alkalmazó algoritmus tulajdonképpen két – különböző algoritmussal történő – fordítási lépésnek veti alá a szöveget (egy kézi a plagizáló által és egy gépi az ellenőrzéskor), majd az ezek után kapott, visszafordított szöveget hasonlítja össze az eredeti szöveggel. Esetleg egy adott szöveget kétszer fordít le egy másik nyelvre (egyszer kézzel, egyszer géppel), majd ezeket hasonlítja össze. Mivel a legtöbb mondatnak nincsen egy adott jó fordítása, hanem számos lehetséges fordítása van, így majdnem teljesen biztosak lehetünk benne, hogy komoly különbségek lesznek a mondatok között, nemcsak a szórendben, hanem a használt szavakban, kifejezésekben is. Fischer Márta ezt így fogalmazza meg:

„A nyelvészeti fordítástudomány eredményei – amelynek fontos területe az ekvivalencia kutatása – eloszlathatják azt a téves elképzelést, mely szerint a fordítás automatikus és teljes megfeleltetést (ekvivalenciát) feltételez a két nyelv között. A kutatók különböző megközelítései és a számtalan ekvivalencia-elmélet éppen arra világítanak rá, hogy az ekvivalencia több szinten, több szempont szerint értelmezhető Ezek ismerete tehát éppen abban erősítheti meg a tanulót, hogy nincs egyetlen helyes (ekvivalens) válasz.” [5]

Magyar nyelv esetében további hátrány, hogy a gépi fordítók igen rosszak, a legjobb angol-magyar nyelvpár esetében is tulajdonképpen majdnem minden mondatban hibáznak, és minél összetettebb a mondat, annál valószínűbb, hogy teljesen félre is fordítanak valamit.

Angol-német nyelvpár esetén már el lehet talán gondolkodni, hogy egy automatikus fordító alapján készítsünk egy algoritmust, de még ott is számos hiba adódik. Emellett komoly hátrány, hogy egy külső programra vagy algoritmusra kell hagyatkozni, hiszen a jó minőségű algoritmusok mind fizetősek, így nagyobb mennyiségű szöveg rendszeres lefordítása komoly költségekkel is járna. A Google Translate meghívható egy API-n keresztül, és korábban lehetett is nagyobb mennyiségű szöveget fordítani rajta, de pár hónapja a Google úgy döntött, hogy még fizetség ellenében sem engedi napi 100 000 karakternél nagyobb szöveg lefordítását. Ez még egy rövidebb diploma ellenőrzéséhez is kevés.

„The Google Translate API has been officially deprecated as of May 26, 2011. We are not currently able to offer additional quota.”

2.1. Az algoritmus kialakítása

Két nyelv között a legkisebb egyezés egy **szó** egyezése lehet. Természetesen, ha egy angol szövegben az *eleven* szót olvashatjuk, akkor annak magyarul nem az *eleven* szó fog megfelelni, hanem a *tizenegy* vagy a *11*, de ennek ellenére beszélhetünk egyezésről. Ugyanakkor érdemes megjegyezni, hogy számos szónak nem lesz megfelelője a másik nyelvben, vagy egyáltalán nem is lesz megfelelője, vagy nem szóként jelentkezik. Most a teljesség igénye nélkül vegyünk sorra pár lehetséges eltérést.

- Összetett szavak: elképzelhető, hogy míg az egyik nyelvben egy gondolatot egy szóval, addig a másikban több szóval fejezünk ki, mint például *tavaly és last year*. Fordítva pedig, míg magyarul *szabadlábra helyeznek* valakit, angolul ezt a jelentést a *liberated* adja vissza.

- Ragozás: a magyar nyelv (akárcsak például a török) számos dolgot ragokkal, a szóval egybe írva fejez ki, míg más nyelvek erre elöljárót használnak. Ami magyarul az *álmomban*, az angolul *in my dream* történt.
- Antoníma: gyakran egy kifejezést jobb antonímával fordítani, nem önmagával. Míg magyarul valami *nem felel meg a célnak*, addig ugyanez angolul *inadequate*.
- Ismétlések elkerülése: bizonyos nyelvek, mint például a magyar, kevésbé szeretik az ismétlést, és inkább utalnak az ismétlődő dolgokra, illetve szinonimákat használnak. A „80 nap alatt a föld körül" magyar fordításában találkozunk a *gentleman* szóval, ahol az angolban a *Mr. Fogg* szerepel.
- Teljes átalakítás: kifejezések és a forrás- valamint célnyelv különbözőségén, illetve a két olvasótábor kulturális ismeretének a különbözőségéből adódóan. A *Queen's pudding*-ból *rakott palacsinta* lesz, az *egg and spoon races* pedig *ügyességi gyerekjáték*. [6]

Azaz számos eset képzelhető el, amikor egy adott szó nem felel meg egyértelműen a másik nyelv egy szavának, ugyanakkor a szavak jelentős része megtalálható lesz mindkét nyelvben. Ugyan a szavakat jól fel lehet használni arra, hogy fordításokat keressünk, de önmagában két szöveg még nem lesz azonos pusztán azért, mert sok közös szavuk van.

Ha eggyel magasabb szintre lépünk, a **tagmondat**ok szintjére, akkor azt látjuk, hogy bár gyakran előfordul a tagmondatok egyezése, de míg a magyarban igen sok vesszőt használunk, és legtöbbször egyértelműen jelöljük a tagmondatok határát, addig az angol nyelvben alig vannak vesszők, és kimondottan nehéz feladat a tagmondatok határának megkeresése. Emiatt ezzel a lehetőséggel most itt nem is foglalkozunk.

A következő szint a **mondat**ok szintje. Ha valaki nekiáll egy szöveg fordításának, akkor azt az esetek túlnyomó részében mondatonként fordítja le. Egy irodalmi fordítás esetén gyakrabban találkozunk azzal, hogy egy mondatot kettőbe szed a fordító, vagy két mondatot összevon, de még itt is viszonylag ritkán fordul elő ez a gyakorlat.

Az ennél magasabb szintekkel, **bekezdés**ekkel, **fejezet**ekkel ugyanaz a legnagyobb gond, mint a tagmondatokkal: nem egyértelmű a jelölésük, elhagyhatóak, összevonhatóak, így ezek egyezésének a vizsgálatára úgyszintén nem térünk most ki.

Mint láttuk, fordítások esetében a legértelmesebb szint a szavak vagy a mondatok szintje. A szavak esetében viszont lényeges a szó többi szóhoz viszonyított pozíciója, a szövegkörnyezet, hiszen bármely két azonos nyelven íródott szövegben vannak azonos szavak, még akár ezek mértéke is magas lehet, azonban ekkor sem biztos, hogy a két szövegnek ugyanaz a jelentése, vagy esetleg csak a témája egyezik. Mint azt a webes keresők esetében látjuk – ahol adott szavakat tartalmazó szövegekre keresünk – nagyon nagy az olyan találatok száma, amelyek ugyan megfelelnek a keresőkérdésnek, de semmi közük sincs ahhoz, amit kerestünk. Azaz önmagában a szavak egyezősége nem tesz két szöveget egymás másolatává, nem lehet általa megállapítani a plagizálás tényét. Ez két különböző nyelv esetében még inkább így lesz, hiszen egy adott szónak a másik nyelvben számos másik felel, vagy felelhet meg, így még ez is komoly bizonytalanságot eredményez.

Természetesen ez nem azt jelenti, hogy a szavak nem használhatók két szöveg közti egyezés megtalálására, de önmagában ez nem elég: hiszen ha valaki lefordít egy egyoldalas szöveget angolról, és beteszi a 120 oldalas magyar diplomájába, akkor ennek a megtalálása csak a szavak használatával lehetetlen. Mindenképpen definiálnunk kell egy szövegkörnyezetet, ahol a szavakat keressük. Ezért a kutatáshoz a legjobb kiindulási pontnak a mondatalapú keresés tűnt, ahol a szavaknak van szövegkörnyezetük (egy mondat), ráadásul a mondat már elég egyedi ahhoz, hogy két dokumentumban – még ha azonos témában íródtak is – nagyon kicsi annak az esélye, hogy két azonos mondat lesz (rövid, egy-, két-, háromszavas mondatokat és közös idézeteket nem számítva). Könnyen beláthatjuk ezt, ha belegondolunk, hogy a legtöbb nyelvnek több százezer szava van [7], a nyelvtani szabályokat most figyelmen kívül hagyva, százezer szóval számolva az adott nyelven egy *n* szóból álló mondat (S_n) összes lehetséges változata:

$$|S_n| = (2 \cdot 10^5)^n$$

Ez egy még hosszúnak sem mondható 10 szavas mondat esetében:

$$|S_{10}| \approx 10^{53}$$

Természetesen ennek a jelentős része értelmetlen mondatot eredményezne, de ennek a hatalmas számnak még az egy tízezreléke is hatalmas. Ha hozzávesszük, hogy például a magyar nyelvben a legtöbb szónak számos alakja van, akkor ez a szám még jelentősen növekedne, de az angol nyelv esetében is a többesszám és egyéb alakok miatt az alapszókincs többszöröse a ténylegesen előforduló szóalakok száma. Ezért tekinthetünk úgy egy mondatra, mint egyedi alkotásra. Sokak szerint egy mondatnál kezdődik a plagizálás, azaz egy (tartalmas, hosszabb) mondat már rendelkezik annyi egyedi tulajdonsággal, hogy lemásolása esetén lehet plagizálásról beszélni.

Érdemes megnézni a Wikipédia ide vonatkozó oldalán található összefoglaló táblázatot, amelyből itt csak egy kivonatot mutatunk be. [8]

Dokumentum, bemeneti adat, szövegkörnyezet	**Szavak száma**	$\lvert S_{10} \rvert$
Egy szöveg leggyakoribb szavai közül ennyi adja ki annak 25%-át.	15	5,8E+11
Egy szöveg leggyakoribb szavai közül ennyi adja ki annak 60%-át.	100	1,0E+20
Kb. egy 2 éves gyerek szókincse	300	5,9E+24
Az Ogden-féle egyszerű angol nyelv (Basic English) szókincse	850	2,0E+29
Ennyi szót használnak az első osztályosok olvasástanításában.	1000	1,0E+30
Kb. egy 6 éves gyerek szókincse	2500	9,5E+33
Arany János Toldi c. művében felhasznált szókincse	3000	5,9E+34
Az átlagember aktív szókincse (élő-aktív és szunnyadó-aktív)	3 000-5 000	5,9E+34

Középfokú nyelvtudásnak megfelelő szókincs	3 500-3 900	2,8E+35
Kb. egy 11 éves gyerek szókincse	5 000	9,8E+36
Az átlagember passzív szókincse	5 000-10 000	5,6E+38
Ennyi szóval a Shreket 95%-ban megértjük.	6 000	6,0E+37
Ennyi szó szükséges a 20. századi angol próza megértéséhez.	8-9 000	1,1E+39
Ennyi szóval a tankönyveket 95%-ban megértjük.	10-12 000	1,0E+40
Egy kétnyelvű kisszótár terjedelme (címszavak)	10-30 000	1,0E+43
Shakespeare (műveiben felhasznált) szókincsét ennyire becsülik	18-25 000	1,7E+43
Petőfi Sándor verseiből kimutatható szókincse	22 719	3,7E+43
Egy átlag értelmiségi egyévi beszédét gondolatban rögzítve kb. ennyiféle szó fordulna elő.	25-30 000	3,0E+44
Igen művelt embereknél a passzív szókincs nagysága	50-60 000	2,5E+47
Kb. ennyi mai magyar szót tartanak számon.	60-100 000	1,1E+49
Egy kétnyelvű nagyszótár terjedelme (címszavak)	120 000	6,2E+50
A 20 kötetes Oxford English Dictionary 2. (nyomtatott) kiadásából (1989) a ma is használt szavak száma	171 476	2,2E+52
A 20 kötetes Oxford English Dictionary 2. (nyomtatott) kiadásának (1989) terjedelme (címszavak)	291 500	4,4E+54
A 33 kötetes Deutsches Wörterbuch terjedelme (1960-as kiadás, címszavak)	350 000	2,8E+55
A Webster's Third New International Dictionary, Unabridged terjedelme (címszavak)	>450 000	3,4E+56
A magyar nyelvben kb. ennyi szó (lexéma!) van (túlnyomórészt elavult vagy rendkívül speciális szavak)	1 000 000	1,0E+60
Az 1,48 milliárd szövegszót (v. szóelőfordulást) tartalmazó magyar webkorpusz 4%-os hibatűréssel készült metszetéből kinyert szókincs mérete (lexémák, ill. szótári szavak), kézi ellenőrzés nélkül	7 200 000	3,7E+68

Jól látható a táblázatból, hogy már egy kétéves gyerek is több száz szót ismer, és ha csak a rövidebb mondatokat vesszük, akkor is több tízezer mondatot tud elméletileg összetenni.

Összefoglalva az előzőeket, láthatólag a mondat egy értelmes egységnek tűnik ahhoz, hogy plágiumot, illetve szövegek közötti egyezéseket keressünk. Ennek az alábbi előnyei vannak:

- Egy értelmes gondolati egységet képvisel
- A mondathatárok nagy pontossággal meghatározhatóak
- A mondat elég egyedi ahhoz, hogy két szöveg között több mondat egyezésekor már valami közös forrást feltételezzünk
- Fordítások esetén a mondat a fordítás egysége, amely mint egység legtöbbször megmarad a különböző nyelvek között [9]

- Egy mondat és fordítása között ekvivalencia van, amely biztosítja, hogy a két mondat jelentése minél közelebb legyen egymáshoz

Miután beláttuk, érdemes a mondatok közötti hasonlóságot vizsgálnunk ahhoz, hogy a fordítási plágiumot megtaláljuk, definiálnunk kell egy metrikát, amely a különböző nyelven íródott mondatok közötti hasonlóság mértékét határozza meg.

2.2 A hasonlósági metrika

Mint korábban említettük, egy angol és egy magyar nyelvű mondat szavai – ha nem is teljes mértékben –, de megfeleltethetők egymásnak. A két nyelv nyelvtanának különbségéből és a magyar nyelv kötetlen szórendjéből adódóan a szavak sorrendje teljesen lényegtelen ebben a megfeleltetésben, azaz az angol nyelvű mondat első, második, harmadik… szava bárhol lehet a magyar mondatban, és fordítva.

A sorrendet figyelembe nem vevő, egy szöveg szavait reprezentáló modell a szózsák (bag of words) [10] – egy adott szöveg összes szavát tartalmazó, de a sorrendet figyelembe nem vevő halmaz –, amelyet számos helyen használnak a szakirodalomban például dokumentumok csoportosítására, spamszűrésre, de még érzelmek felismerésére is [11]. Mi most sokkal kisebb egységben, a mondatok szintjén fogjuk a szózsákot alkalmazni.

Egy *n* szóból álló mondatot (*S*) képviseljenek a benne lévő szavak (*w*).

$$w_x \in S_x \text{ és } w_y \in S_y$$

Természetesen ez egy egyszerűsítés, hiszen elméletileg ugyanazokból a szavakból más mondatokat is össze lehet rakni. Azonban, mivel az esetek túlnyomó részében elég egyértelműen visszaállítható a mondat értelme a szavak ismeretében, túl sok hibát ez az átalakítás nem fog eredményezni.

$$S_x = \{w_{x1}, w_{x2}, w_{x3}, \ldots w_{xn}\}$$

Most definiáljuk két mondat hasonlóságának a mértékét (*Sim*) a bennük levő közös szavak számával.

$$Sim(x,y) = | S_y \cap S_z |$$

Ez már egy jó megközelítés, de számos dolgot nem vesz figyelembe. Például egy hosszú és egy rövid mondat hasonlósága így maximum akkora lehet, amekkora a rövid mondat hossza. Ez helyes is, ugyanakkor például ha a hosszú mondatban megtalálható a rövid mondat összes szava, akkor ez a két mondat ugyanannyira hasonló lesz, mintha a rövid mondatot önmagával hasonlítottam volna össze, ami viszont egyértelműen rossz: ezért figyelembe kell venni nemcsak a közös szavakat, hanem a hiányzó szavakat is. Ezeket érdemes súlyozni is, most legyen a megtalált szavak súlya α , a nem megtaláltaké β.

$$Sim(x,y) = \alpha \cdot | S_x \cap S_y | - \beta \cdot | S_x \setminus S_y |$$

Amennyiben α értékét 3-nak, β értékét pedig 1-nek vesszük, akkor az azt jelenti, hogy minden olyan szót, amelyik megvan a másik mondatban, háromszoros súllyal veszünk figyelembe a hiányzó szavakhoz képest.

Ez a képlet már majdnem tökéletes, de nem szimmetrikus $S_x \setminus S_y$ miatt, azaz: $Sim(x,y) \neq Sim(y,x)$. Ez nem jó így, hiszen annak az esélye, hogy S_x S_y-nak a fordítása elvileg ugyanannyi kell legyen, mint annak esélye, hogy S_y S_x-nek a fordítása. Ezt a hibát úgy lehet kiküszöbölni, hogy például kiszámoljuk mindkét értéket, majd ennek vesszük az összegét. Ugyanakkor azért vezettük be az egyenlet második tagját ($S_x \setminus S_y$), mert azok a szavak, amelyek csak az egyik mondatban találhatóak meg, csökkentik annak valószínűségét, hogy a két mondat egymás fordítása. Ha annak az esélye, hogy S_x fordítása S_y-nak kisebb, mint a fordítottja azaz $Sim(x,y) < Sim(y,x)$, akkor ez a legtöbb esetben azt jelenti, hogy S_x hosszabb, azaz több olyan szó van benne, aminek nincs fordítása a másik mondatban. Ez lényeges: hiába kapunk Sim(y,x)-re egy nagyon magas értéket, ha Sim(x,y) alacsony, hiszen akkor majdnem biztos, hogy a két mondat nem fordítása egymásnak, esetleg az egyik a másik része. Ezért a továbbiakban úgy számoljuk ki Sim(x,y) értékét, hogy a korábban definiált értékek közül az alacsonyabbat vesszük. Ezzel az új képlet:

$$Sim(x,y) = \min (\alpha \cdot | S_x \cap S_y | - \beta \cdot | S_x \setminus S_y | ,$$
$$\alpha \cdot | S_y \cap S_x | - \beta \cdot | S_y \setminus S_x |)$$

Ez a definíció már eleget tesz a szimmetria (ekvivalencia) követelményének, azaz most már

$$Sim(x,y) = Sim(y,x)$$

A továbbiakban még néhány lényeges dolgot figyelembe kell vennünk ahhoz, hogy a szózsák algoritmus fordítások esetében is jól működjön. Mivel S_x és S_y nyelve nem azonos, ezért definiálnunk kell, hogy mit jelent két szó azonossága, illetve különbözősége: azaz mikor mondjuk, hogy $w_x \equiv w_y$ és mikor mondjuk, hogy $w_x \not\equiv w_y$ Ahhoz, hogy ezt meghatározzuk, definiálnunk kell még egy műveletet, a fordítás műveletét, azaz egy fordítási függvényt, amely egy szónak, illetve annak összes szótövének az összes fordítását adja vissza a másik nyelven.

$$trans(w_x) = W_y \text{ ahol } w_y \in W_y$$

$$trans(w_y) = W_x \text{ ahol } w_x \in W_x$$

mivel a fordítás egy szimmetrikus művelet, ezért ha

$$w_x \in trans(w_y) \text{ akkor } w_y \in trans(w_x)$$

ezek alapján definiáljuk, ha

$$w_y \in trans(w_x) \text{ akkor } w_x \equiv w_y$$

illetve ha

$$w_x \in \text{trans}(w_y) \text{ akkor } w_x \equiv w_y$$

hasonló módon ha

$$w_y \notin \text{trans}(w_x) \text{ akkor } w_x \not\equiv w_y$$

illetve ha

$$w_x \notin \text{trans}(w_y) \text{ akkor } w_x \not\equiv w_y$$

A fent leírt algoritmusnak számos előnye van: először is nem kell szóegyértelműsítést használni, hiszen az azonossági függvényünk – amelynek pontos működésének leírásától eltekintünk, csak a definícióját adtuk meg – ezt feleslegessé teszi azzal, hogy minden lehetséges jelentést figyelembe vesz. Az egynyelvű plágiumkeresésekben használt szinonima-egyértelműsítést, illetve -szűrést sem kell alkalmazni, hiszen egy szónak a lehetséges fordításai a másik nyelven egy vagy több szinonimahalmazba rendezhetőek, és ezeket az algoritmus transzparensen kezeli. Az algoritmus nem érzékeny a szavak sorrendjére, mint az n-gram algoritmus, azaz nem függ a fordítástól és nem működik nagyon eltérően barátságos és nem barátságos nyelvpárok esetében. Az algoritmus hátránya viszont a hatalmas keresési tér és a lineáris keresési idő, azaz a keresés ideje lineárisan függ az adatbázis méretétől. Nagy adatbázisok esetén ez gyorsan elfogadhatatlan keresési időket eredményez. Ez utóbbi problémát az implementációs fázisban egy indexált kereséssel meg tudtuk oldani, de most a részletek ismertetésétől – helyszűke miatt – eltekintünk.

2.3. Tesztkörnyezet kialakítása

Az algoritmus teszteléséhez szükségünk van olyan szövegekre, amelyeknek ismerjük a fordítását, valamint egy olyan hatalmas korpuszra, amely lehetővé teszi a hamis pozitív találatok tesztelését is, azaz egy olyan korpuszra, amely már biztos tartalmaz hasonló mondatokat, hiszen 10 mondatból kiválasztani egy adott mondat fordítását egy igen rosszul teljesítő algoritmusnak se lenne gond. Nagyméretű korpusznak a Wikipédiát választottuk, abból is az angol nyelvűt. [12] Amennyiben egy algoritmus képes egy Wikipédia méretű adatbázisból kiválasztani a megfelelő mondato(ka)t, akkor elmondhatjuk, hogy jól működik. Utóbbira azért is esett a választás, mert sokan idéznek, illetve sokan plagizálnak is sajnos a Wikipédiából, így gyakorlati haszna is van egy olyan keresőnek, amely kiemeli a Wikipédiából átvett részeket egy dolgozatban. Szótövezésre a MOKK által fejlesztett, ingyenesen elérhető Hunspellt alkalmaztuk [13]. Számos eszköz létezik, amely képes szövegeket mondatokra bontani, de mi három okból döntöttünk a saját algoritmus használata mellett: a) Először is a Wikipédia szövege – még szöveges formátumra alakítás után is – tartalmazott hibákat, például mondatok rendszeresen egybeíródnak a következővel (hiányzik a szóköz a mondatot lezáró írásjel után). b) Másodszor pedig egy olyan algoritmusra volt szük-

ségünk, ami gyors, és segítségével elkerülhetjük az újabb köztes fájlok létrehozását. c) Mivel ekkor már látszott, hogy a teljes folyamat igen erőforrás-igényes, ezért szerettünk volna minél kevesebb külső programot használni, hogy a plágiumkereső program minél több gépen legyen képes futni.

Több okból kifolyólag is elengedhetetlennek bizonyult egy automatikus fordító használata a tesztekhez. Az első és legfontosabb, hogy nem rendelkezünk annyi Wikipédiából – vagyis tulajdonképpen bárhonnan – származó angol-magyar párhuzamos korpusszal, amely elegendő lenne az algoritmus tesztelésére. Természetesen össze kell vetni az automatikus fordítóval és egy személy által fordított szövegen elért eredményeket, hogy megbizonyosodjunk arról, hasonló eredményt kapunk a két esetben. A könnyű elérhetőség és az API felület miatt esett a választás a Google fordítójára. [14]

Ahhoz, hogy egy angol és egy magyar szó azonosságát meg tudjuk állapítani, szükségünk van egy szószedetre, egy lapos szótárra. Ehhez kitűnő alapot nyújtott a SZTAKI online szótára. [15] Mivel azt is szükséges tesztelni, hogy a szótár mérete, illetve a hiányzó fordítások mennyire befolyásolják az algoritmust, ezért más, online elérhető szótárakkal illetve szószedetekkel is végeztünk kísérleteket. A kutatás jelentős részét az összes szótár uniójával végeztük.

3 Konklúzió

Az algoritmus teszteléséhez a teljes feldolgozott angol Wikipédiát feltöltöttük egy adatbázisba, és ebben kerestünk, mind a kézzel magyarra fordított, mind a géppel fordított Wikipédia cikkeket. A két keresés között statisztikai különbséget nem találtunk, így most a sokkal nagyobb mennyiségű, géppel fordított korpuszon elért eredményeket ismertetjük.

A magyar mondatokra keresve 0,67 recall értéket kaptunk, azaz ennyi volt az aránya azon mondatoknak, ahol a teljes Wikipédiából sikerült kiválasztanunk azt a mondatot, amelyiknek ez a magyar mondat a fordítása. Ez annyit jelent, hogy egyenletes valószínűséget feltételezve a mondatoknál annak az esélye, hogy egy 10 mondatból álló szakaszból egy hasonlót se találunk meg, 0,000016; és csak az esetek 2%-ban fogunk kevesebb mint 4 mondatot hasonlónak találni.

A recall értéke könnyedén mérhető, amennyiben tudjuk, hogy mit fordítottunk le a másik nyelvre. Ugyanakkor a pontosság meghatározása sokkal körülményesebb, hiszen kézzel kell ellenőrizni, hogy a visszaadott találatok közül melyek tényleges lehetséges fordítások, és melyek nem. Egy véletlen kiválasztott, kézzel fordított, és kézzel ellenőrzött korpusz esetében, ahol α értékét 2-nek, β-t pedig 1-nek választottuk, a hasonlósági metrika (*Sim*) minimumát pedig 8-nak, a pontosságra 0,92-t kaptunk, a recall értéke pedig 0,85 lett. Ebből F_1=0,88 adódik.

Az algoritmus kutatása már befejeződött, jelenleg az algoritmus finomhangolásán és a KOPI Plágiumkereső Portálba való integrálásán dolgozunk. A konferenciára már mindkettő elkészül és reményeink szerint be tudunk számolni az első publikus tesztek eredményéről is.

4 További tervek

Az algoritmust kézzel ellenőriztük más nyelvpárok esetében is, és az eredmények bíztatóak, de célunk, hogy pontosan kiszámoljuk a recall és pontosság értékeket legalább 10 további nyelvpár esetében is.

A szószedet mérete lineáris összefüggést mutat a futási idővel, azaz minél több lehetséges fordítása van egy szónak, annál nagyobb a keresési tér, és annál lassabb lesz a keresés. A pontosságot ugyanakkor sokkal kisebb mértékben javítja egy adott mérethatár felett, így meg kell határozni, hogy mi az ideális szószedet mérete, amely még gyors algoritmust eredményez, de már a találati pontossága is megfelel egy adott alkalmazáshoz. Ez a méret valószínűleg nyelvpárfüggő lesz.

Az algoritmus működik egynyelvű keresések esetében is, amennyiben a fordítási azonosság (*trans*) helyett szinonimákat, antonimákat, hiper- és hiponimákat használunk. Össze szeretnénk hasonlítani az egynyelvű keresést a jelenleg legtöbb plágiumkereső által használt n-gram algoritmus eredményével is.

Bibliográfia

1. Bailey, J: The Problem with Detecting Translated Plagiarism, http://www.plagiarismtoday.com/2011/02/24/the-problem-with-detecting-translated-plagiarism/ (2011)
2. Dr. Weber-Wulff, D.: Results of the Plagiarism Detection System Test 2010, http://plagiat.htw-berlin.de/software-en/2010-2/ (2010)
3. PAN 2010 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse http://www.uni-weimar.de/medien/webis/research/events/pan-10/ (2010)
4. Potthast, M.; Barrón-Cedeño, A.; Eiselt, A.; Stein, B.; Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection, http://www.clef2010.org/resources/proceedings/clef2010labs_submission_125.pdf (2010)
5. Fischer, M.: Fordítás és közvetítés a nyelvoktatásban – mit nyújthat a nyelvoktatásnak a fordítástudomány? , http://ecml.opkm.hu/files/FischerM.doc (2008)
6. Tóth, P.: Fordításelmélet, http://dettk.ucoz.com/load/0-0-0-93-20 (2005)
7. How many words are there in the English language?, Oxford University Press, http://oxforddictionaries.com/page/93 (2011)
8. Wikipedia, Szókincsméretek összehasonlító listája, http://hu.wikipedia.org/wiki/Szókincsméretek_összehasonlító_listája (2011)
9. Nida, E. A.: Toward a Science of Translating. E. J. Brill, Leiden (1964)
10. Wikipedia: Bag of words model, http://en.wikipedia.org/wiki/Bag_of_words_model (2011)
11. Miháltz, M.: OpinHu: online szövegek többnyelvű véleményelemzése. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2010)
12. Wikipedia the free encyclopedia, http://en.wikipedia.org/ (2011)
13. BME MOKK: Hunspell szótövező, helyesírás ellenőrző, morfológiai elemző, http://hunspell.sourceforge.net/ (2011)
14. Google: Google Translate, http://translate.google.com/ (2011)
15. MTA SZTAKI: SZTAKI Szótár, http://szotar.sztaki.hu/ (2011)