Minőségbecslő rendszer egynyelvű természetes nyelvi elemzőhöz

Yang Zijian Győző[1], Laki László János[2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter utca 50/a
{yang.zijian.gyozo, laki.laszlo}@itk.ppke.hu

Kivonat A pszicholingvisztikai indíttatású természetes nyelvi elemzés egy új, emberi nyelvelemzést modellező nyelvtechnológiai módszer. Ez a modell egy valós idejű elemző, amelynek párhuzamosan több szála elemzi egyszerre a bemeneten sorban érkező szavakat, kifejezéseket vagy mondatokat. A párhuzamosan futó szálak közül az egyik a minőségbecslő modul, amely menedzseli, szűri a hibás és zajos bemenetet, valamint tájékoztatja a többi szálat a bemenet aktuális minőségéről. A minőségbecslő modul felépítéséhez a gépi fordítás kiértékeléséhez használt minőségbecslés módszerét használtuk. Ahhoz, hogy a minőségbecslő modellünk a természetes nyelvi elemző egyik párhuzamosan futó szálát képezze, ötvöztük az eredeti minőségbecslő rendszert a feladatorientált architektúrával. A kutatásunk során felépítettünk egy feladatorientált minőségbecslő rendszert, amely az egynyelvű szöveg valós idejű minőségének becslésére alkalmas. Az általunk létrehozott rendszer segítségével ∼70%-os pontossággal tudjuk megbecsülni a bemeneti szöveg minőségét. A rendszer az AnaGramma magyar nyelvű elemzőhöz készült, de más nyelvekre is használható.

Kulcsszavak: minőségbecslés, pszicholingvisztika, természetes nyelvi elemzés

1. Bevezetés

Mára a pszicholingvisztika fontos terület lett a számítógépes nyelvészetben. Amíg a hagyományos nyelvi elemzők (pl.: szintaktikai elemzők, szófaji elemzők stb.) a mondat végének elhangzása után kezdik az elemzést, addig az emberi elemző a kommunikáció során folyamatosan dolgozza fel a hallott vagy az olvasott szavakat, kifejezéseket.

Az AnaGramma [3,9] egy pszicholingvisztiaki indíttatású nyelvi elemző rendszer, amely modellálja a valós emberi nyelvi feldolgozást. Az elemző performancia alapú és szigorúan balról jobbra elemzi a bementet. A rendszer architektúrája eredendően párhuzamos. A hagyományos megközelítésekkel szemben, itt az elemzendő szót valós időben, folyamatosan dolgozza fel. A párhuzamosan jelenlévő szálak (pl.: szintaktikai elemző, morfológiai elemző, korpuszgyakorisági szálak stb.) egyszerre és egymással kommunikálva vizsgálják a bemenetet, valamint

egymás hibáit javítva végzik el az elemzést. Ezen párhuzamos szálak közül az egyik fontos szál a minőséget becslő, vizsgáló szál.

A minőségelemző szál legfontosabb feladatai, hogy minőségi jelzőket biztosít a többi szál és a felhasználó számára, valamint adott esetben kontrollálja, szűri a bemenetet.

Az AnaGramma alapvetően kétféle száltípust használ. Az egyik típus a *felkínálás* jellegű szál, amely információt ad az elemről (pl.: alanyesetű), a másik típus pedig az *igény* jellegű szál, amely egy adott tulajdonságú elemet vagy szálat (pl.: a birtok igényel egy alanyesetű vagy datívuszos alakot) keres. A felkínálások és az igények feldolgozása során azonban túl nagy mennyiségű felkínálás keletkezhet egy adott igényhez, amelyek közül számtalan az irreleváns elem. A minőségelemző szál megfelelő specifikus jegyekkel (feature) képes szűrni a felkínálásokat, valamint ha több helyes felkínálás is van, segíthet kiválasztani a megfelelő felkínálási elemet.

A különböző szálaknak különböző minőségi jelzőkre van szüksége, valamint a felhasználót is folyamatosan tájékoztatni kell az aktuális minőségről. Ezért egy olyan minőségbecslő rendszerre van szükség, ami rugalmasan változtatható, bővíthető és ütemezhető. Továbbá a feladatoknak megfelelően a különböző jegyek különválaszthatóak legyenek. A hagyományos minőségbecslő rendszerek, mint a QuEst [12], nem tudják kielégíteni változtatás nélkül ezen követelményeket (más célt szolgálnak), ezért létrehoztunk, a hagyományos minőségbecslő rendszerekre építkezve, egy új minőségbecslő architektúrát.

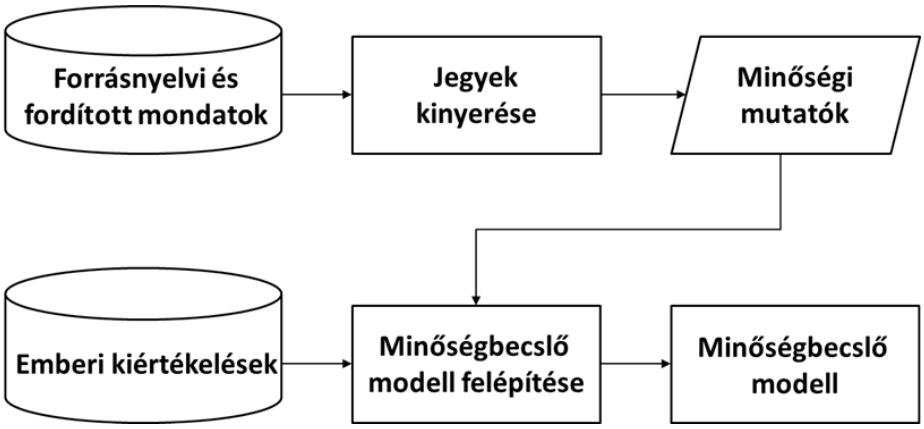
2. Kapcsolódó munkák

A hagyományos minőségbecslő módszer (lásd 1. ábra) különböző minőségi mutatókat nyer ki a forrás és a gép által lefordított mondatokból. Majd gépi tanulással betanítja a minőségi mutatókat az emberek által kiértékelt mutatókra. Az így betanított modell segítségével tudja megbecsülni az új ismeretlen mondatok minőségét. Mivel a gépi tanulás modellje emberi kiértékelésen alapszik, ezért a becsült értékek magasan korrelálnak az emberi kiértékeléssel.

Az AnaGramma egy egynyelvű elemző, ezért a minőségbecslő modellünk tanításához egynyelvű korpuszt használtunk.

A QuEst++ [11] rendszer szó szintű elemző része tartalmaz egynyelvű kiértékeléseket, többek között nyelvi modell jegyeket, szintaktikai jegyeket, célnyelvi kontextus jegyeket stb. De ezek a kiértékelések csupán egy apró részét képezik a gépi fordítást kiértékelő rendszernek és nem egy kifejezetten egynyelvű minőségbecslő rendszer.

Számtalan kutatási és üzleti ágban használatos a feladatorientált vagy szolgáltatásorientált architektúra. Ilyen területek például az elektronikus kereskedelem [7], a robotika [8], az automatikus videó megfigyelő rendszerek [5] stb. A feladatorientált architektúra előnye, hogy a modell feladatokban gondolkozik. Minden feladat egy független egység, amely önálló funkcionalitással bír. A különböző feladatok különböző erősforrásokat, valamint eszközöket használhatnak és akár párhuzamosan egyszerre többféle problémát is megoldhatnak.



1. ábra. A minőségbecslő modell

Hatékony ütemezéssel rugalmasan optimalizálhatjuk a teljesítményt és a különböző specifikus igények kiszolgálását.

A kutatásunkban a hagyományos minőségbecslő rendszert alakítottuk át a feladatorientált architektúrával.

3. πRate rendszer

A kutatásunk során implementáltunk egy egynyelvű minőségbecslő rendszert, a πRate[3] rendszert. Kettő fő modulja van a rendszernek (lásd 2. ábra): a tanuló modul és a kiértékelő modul.

A tanuló modul legfőbb feladata, hogy betanítja a minőségbecslő modellt. Ez a modul megegyezik a hagyományos minőségbecslő modell tanuló moduljával (lásd 1. ábra).

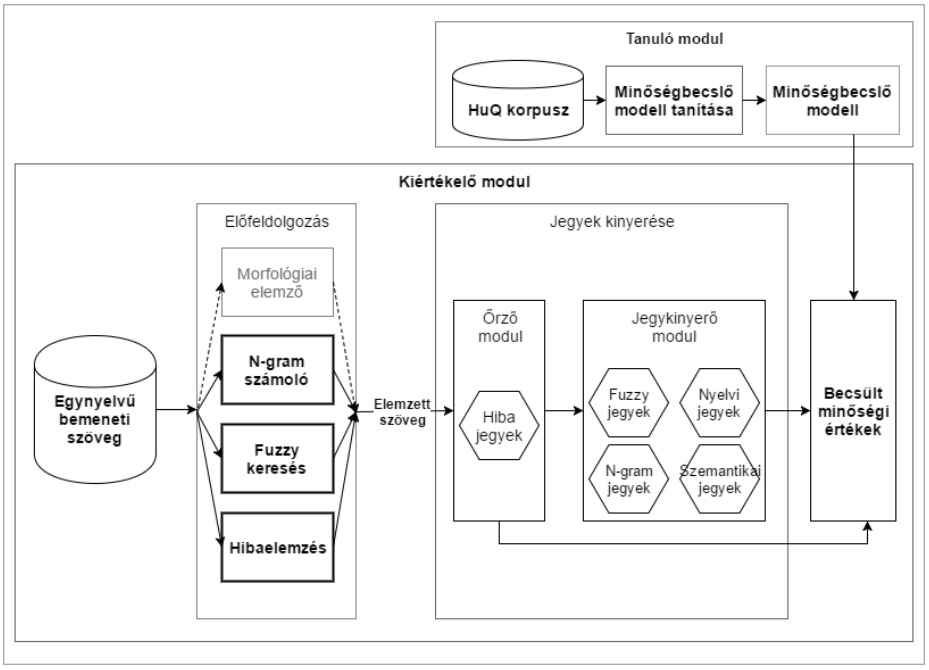
A tanításhoz emberek által kiértékelt egynyelvű korpuszt használtunk. A gépi tanuláshoz különböző nyelvi és statisztikai jegyeket használtunk: pl.: nyelvi jegyek; n-gram jegyek; fuzzy jegyek; hiba jegyek stb.

A korpuszból a jegyek segítségével kinyertük a minőségi mutatókat, majd a mutatók segítségével betanítottuk a minőségbecslő modellt az emberek által kiértékelt minőségi mutatókra.

A másik fontos modul a kiértékelő modul. Első lépésként a kiértékelő modul beolvassa a bemeneti szöveget, majd az előfeldolgozó fázisban elemezzük azt. Az így elemzett szöveget adjuk tovább a jegykinyerő (feature extraction) modulnak, amely a különböző jegyek és a betanított minőségbecslő modell segítségével előállítja a minőségi mutatókat.

A bemenet lehet nyers vagy elemzett szöveg. A feladatorientált πRate rendszer egyik előnye, hogy a morfológiai elemzés műveletét kihagyhatjuk az előfel-

[3] A kutatólaborunk a 314-es szobában található.



2. ábra. A πRate rendszer

dolgozó fázisban (a 2. ábrában a morfológiai elemző szürke), mivel az AnaGramma esetében a minőségbecslő szál már egy morfológiailag elemzett szöveget kap, amellyel így képes optimalizálni az erőforrást és ezáltal a teljesítményt.

A kiértékelő modulnak három fő része van:

– előfeldolgozó modul;
– ellenőrző modul;
– jegykinyerő modul.

A bemeneti szöveg inkrementálisan bővül. Amint a szöveg beérkezik a πRate rendszerbe, az előfeldolgozó modul elemzi a szöveget például morfológiailag elemzi (ha az még nem elemzett) a szöveget, kiszámolja az n-gram valószínűségeket stb. Majd az elemzett szöveget továbbküldi a jegykinyerő modul számára, ahol először az ellenőrző modul a hiba jegyek segítségével leellenőrzi a szöveget. Ha a szöveg hibamértéke meghalad egy megadott küszöbértéket, akkor az ellenőrző modul felhatalmazást kaphat megszakítani a folyamatot vagy szűrni, "cenzúrázni" a szöveget. Máskülönben továbbengedi a többi jegyek számára a szöveget és a saját hibaértékeit minőségi mutatóként használja fel a minőségbecslő modell.

A jegyek kinyerése után a minőségi mutatókkal a minőségbecslő modell kiszámolja a becsült értékeket (pl.: szálspecifikus értékek, aktuális mondat minősége, eddig beolvasott összes szöveg globális minősége stb.).

4. Módszerek és mérések

A minőségbecslő modell felépítéséhez egynyelvű jegyekre van szükségünk, amelyek segítségével kinyerjük a minőségi mutatókat. A tanítás során a jegyek egy egynyelvű korpuszból nyerik ki a szükséges értékeket. Majd gépi tanulással emberek által kiértékelt minőségi mutatókra tanítjuk be a modellt (lásd 1. ábra). A πRate rendszer felépítéséhez JAVA EE-t használtunk.

4.1. HuQ corpus

A minőségbecslő modell tanításához a HuQ korpuszt [14] használtuk. A HuQ korpusz 1500 magyar mondatot tartalmaz. Mind az 1500 mondatot három magyar anyanyelvű ember értékelte ki. A kiértékeléshez a Likert értékskálát használták, ebben az esetben 1-től 5-ig lehetett értékelni a mondatok minőségét. A HuQ korpusz továbbá tartalmaz még osztályzási értékeket is: BAD: 1 <= minőség <=2; MEDIUM: 2 < minőség < 4; GOOD: 4 <= minőség <=5. A korpusz vegyes témájú: film feliratok, irodalom és jog.

A kísérletünkben a HuQ korpuszt felosztottuk 500 mondatot fuzzy referenciának és 1000 mondatot minőségbecsléshez.

Az 500 mondatos fuzzy referenciakorpuszt kézzel állítottuk össze. Közel egyenlő arányban tartalmaz "BAD", "MEDIUM" és "GOOD" osztályzatú mondatokat (167 "BAD" mondat, 166 "MEDIUM" mondat, 167 "GOOD" mondat).

Az 1000 mondatot, amelyeket a minőségbecslő modellhez tettünk félre, további 90-10% arányba osztottuk fel tanító és tesztelő halmazra. A teszteléshez tízszeres keresztvalidálást használtunk.

4.2. Egynyelvű jegyek

A minőségbecslő modellünk 32 különböző típusú jegyet használ, amelyeket jellegük alapján az alábbi kategóriákba soroltuk:

- nyelvi jegyek:
  - főnevek, igék, igekötők, melléknevek, határozószók, kötőszók, névmások, névelők, indulatszók aránya a mondatban;
  - főnevek és igék aránya a mondatban;
  - főnevek és melléknevek aránya a mondatban;
  - igék és igekötők aránya a mondatban;
  - főnevek és névelők aránya a mondatban;
- n-gram jegyek:
  - a mondat nyelvmodell valószínűsége;
  - a mondat nyelvmodell perplexitása;
  - a mondat nyelvmodell perplexitása mondatvégi írásjel nélkül;
  - a mondat szótöveinek, szófaji címkéinek nyelvmodell valószínűsége;
  - a mondat szótöveinek, szófaji címkéinek nyelvmodell perplexitása;
- fuzzy jegyek:

  - A Hanna Bechara és társainak szemantikai hasonlóság kutatása [1] alapján a HuQ korpusz egyharmadát referenciakorpuszként használtuk fel. A referenciamondatok közül fuzzy kereséssel megkerestük a bemeneti szöveghez legjobban hasonlító mondatot. Majd a megtalált referenciamondathoz tartozó minőségi értékeket (Likert és osztályzási értékei) a minőségbecslő modellünkben felhasználtuk minőségi mutatóként (jegyként). A fuzzy kereséshez használtuk a Levenstein távolságot, a TER (Translation Error Rate) mértéket, a BLEU mértéket, a NIST mértéket és a szemantikai hasonlóságot mérő LSI [4] módszert és a beágyazási modelleket [10].
- hiba jegyek:
  - xml címkék aránya a mondatban;
  - nem magyar szavak aránya a mondatban;
  - ismeretlen szavak aránya a mondatban;
  - írásjelek aránya a mondatban.

Az egynyelvű jegyek és a HuQ korpusz segítségével felépítettük a minőségbecslő modelljeinket:

- LS modell: minőségbecslő modell Likert értékeket felhasználva.
- OS modell: minőségbecslő modell osztályzási értékeket felhasználva.

4.3. Mérések

A minőségbecslő modell felépítéséhez több gépi tanuló algoritmust is kipróbáltunk, amelyek közül szupport vektor regresszió adta a legjobb eredményeket, ezért a továbbiakban ezt használtuk.

Miután betanítottuk a minőségbecslő modellt, implementáltuk a πRate rendszert. A πRate rendszer előfeldolgozó fázisában: az elemzéshez használtuk a PurePos 2.0 [6] szófaji elemzőt; az n-gram számoláshoz a SRILM [13] eszközkészletet; a fuzzy kereséshez a BLEU, NIST és Levenstein mértékeket; a szemantikai hasonlóság méréséhez az LSI és a beágyazási modelleket. A kiértékelő fázisban az ellenőrző jegyek szűrték a hibás bemenetet, majd a jegykinyerő modul kiszámolta a minőségi mutatókat.

A πRate rendszer jegyhalmazára optimalizálást végeztünk el, úgy ahogyan a hagyományos minőségbecslés módszerében optimalizálni lehet a jegyek halmazát [2].

Az optimális jegyhalmaz megtalálásához a „forward selection” [15] módszert használtuk:

- OptLS halmaz: optimalizált jegyhalmaz LS modellhez.
- OptOS halmaz: optimalizált jegyhalmaz OS modellhez.

5. Eredmények és kiértékelések

A πRate rendszer kiértékeléséhez a MAE (mean absolute error - átlagos abszolút eltérés), az RMSE (root mean square error - átlagos négyzetes eltérés gyöke),

a Pearson-féle korreláció és a helyesen osztályozott egyed (Correctly Classified Instances - CCI) mértékeket használtuk.

A HuQ korpusz és a 32 jegy segítségével betanítottuk a minőségbecslő modellt és felépítettük a πRate rendszert. Az 1. táblázatban és a 2. táblázatban láthatjuk, hogy a 32 jegyhalmazzal ∼59%-os korrelációt és ∼70% helyesen osztályozott egyedet értünk el.

	Correlation	MAE	RMSE
LS modell - 32 jegy	0,5936	0,6857	0,8961
OptLS halmaz - 13 jegy	0,6278	0,6783	0,8758

1. táblázat. LS modell és OptLS halmaz kiértékelése

	CCI	MAE	RMSE
OS modell - 32 jegy	70,7%	0,2465	0,3590
OptOS halmaz - 8 jegy	71,7%	0,2544	0,3539

2. táblázat. OS modell és OptOS halmaz kiértékelése

Az optimalizálciót a „forward selection” módszerrel végeztük el. Ezt szintén láthatjuk az 1. táblázatban és a 2. táblázatban:

– Az OptLS halmaz, 13 jeggyel ∼3%-al magasabb korrelációt tudott elérni.
– Az OptOS halmaz, 8 jeggyel ∼1%-al több helyes egyedet osztályozott.

A 3. táblázatban és a 4. táblázatban láthatjuk az optimalizált jegyhalmazokat (az eredmény minőségjavulásának mértéke alapján van sorba rendezve). Általánosságban azt állapíthatjuk meg, hogy a mondatok nyelvmodell valószínűsége és perplexitása igen fontos szempont. Illetve a beágyazási modell jobban teljesített az LSI modellnél, hiszen az optimalizált halmazokba egy LSI jegy sem került bele. Láthatjuk továbbá azt is, hogy a Fuzzy egyezés modellek is előkelő helyezéseket értek el, ami azt jelenti, hogy nagyban befolyásolja az eredményt.

A 5. táblázatban láthatunk helyes és hibás becsléseket, amelyeknél fontos szempont a fuzzy egyezés. Ha a modell talál a referencia korpuszban hasonló mondatot, akkor az általa kínált értékekkel jó becslést kapunk, de ha a fuzzy kereséssel talált mondat nem hasonlít a bemenetre, akkor erősen rontja a becslés minőségét. Ezért véleményünk szerint fontos, hogy a referencia korpusz mérete nagyobb legyen a jelenleginél, valamint változatos mintákat tartalmazzon, vagyis széles skálában tartalmazzon rövid, hosszú, rossz, közepes és jó minőségű mondatokat.

Jegyek
A mondat szófaji címkéinek nyelvmodell valószínűsége
Kötőszavak aránya
Fuzzy egyezés beágyazási modellel - Likert értéke
A mondat szótöveinek nyelvmodell valószínűsége
Főnevek aránya
NIST fuzzy egyezés (beágyazási modellel) - osztályzási értéke
A mondat szófai címkéinek nyelvmodell perplexitása
Melléknevek aránya
Írásjelek aránya
Igék és igekötők aránya
Igekötők aránya
Ismeretlen szavak aránya
Fuzzy egyezés beágyazási modellel - osztályzási értéke

3. táblázat. Optimalizált 13 jegy Likert modell számára

Jegyek
Az mondat nyelvmodell valószínűsége
Az mondat nyelvmodell perplexitása
Kötőszavak aránya
TER fuzzy egyezés beágyazási modellel - osztályzási értéke
Levenstein fuzzy egyezés (beágyazási modellel - Likert értéke
Az mondat (mondatvégi írásjel nélkül) nyelvmodell perplexitása
A mondat szóteveinek nyelvmodell perplexitása
Írásjelek aránya

4. táblázat. Optimalizált 8 jegy az osztályzási modell számára

Likert értékek		Osztályzási értékek		Sentence
Emberi értékelés	Becsült érték	Emberi értékelés	Becsült érték	
4.333	4.559	GOOD	GOOD	Mahmoud eltorzította az arcát. (Mahmoud contorted his face.)
3.667	3.198	MEDIUM	MEDIUM	Megyek az öltönyt. (I am going the suit.)
2	1.683	BAD	BAD	Az elnök a magát a vége felé, a nebraskai. (The president the himself towards the end, the Nebraskan.)
2	3.531	BAD	BAD	A többi súlyos szó, és hidrokarbon létfontosságú. (The other heavy words and hydrocarbon are vital.)
5	2.520	GOOD	GOOD	Senki sem tudja. (Nobody knows.)
2	3.628	BAD	GOOD	Ők soha csinál amit. (They never does what.)

5. táblázat. Példa jó és rossz becslésre

6. Összegzés

Létrehoztuk a feladatorientált πRate minőségbecslő modellt egynyelvű természetes elemzők számára. Mivel a hagyományos minőségbecslő modell nem tudja kiszolgálni megfelelően a pszicholingvisztikai indíttatású inkrementális elemzőt, ezért a hagyományos módszert a feladatorientált architektúrára módosítottuk. Ennek előnye, hogy rugalmasan ütemezhetjük a jegyeket és hatékonyan tudjuk kiszolgálni a különböző típusú bemeneteket, igényeket és feladatokat.

A πRate rendszer tanításához a HuQ korpuszt és 32 darab jegyet használtunk. A jegyhalmazon optimalizálást végeztünk, amellyel kevesebb jeggyel tudtunk magasabb eredményt elérni. A πRate rendszerrel $\sim$60%-os korrelációt és $\sim$70% helyesen osztályozott egyedet értünk el. A πRate rendszer az AnaGramma természetes elemzőhöz készült, de más rendszerekhez is alkalmazható.

A szoftver készen áll arra, hogy az AnaGramma elemzőbe integrálják, de mivel az AnaGramma még nem készült el teljesen, a további méréseket az integrálás után tudjuk csak elvégezni.

Hivatkozások

1. Bechara, H., Escartin, C.P., Orasan, C., Specia, L.: Semantic textual similarity in quality estimation. Baltic Journal of Modern Computing, Vol. 4 (2016), No. 2 pp. 256–268 (2016)
2. Beck, D., Shah, K., Cohn, T., Specia, L.: Shef-lite: When less is more for translation quality estimation. In: Proceedings of the Workshop on Machine Translation (WMT) (2013)

3. Indig, B., Laki, L., Prószéky, G.: Mozaik nyelvmodell az anagramma elemzőhöz. In: XII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 260–270. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged, Hungary (2016)
4. Langlois, D.: Loria system for the wmt15 quality estimation shared task. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 323–329. Association for Computational Linguistics, Lisbon, Portugal (September 2015), http://aclweb.org/anthology/W15-3038
5. Monari, E., Voth, S., Kroschel, K.: An object- and task-oriented architecture for automated video surveillance in distributed sensor networks. In: Proceedings of the 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance. pp. 339–346. AVSS '08, IEEE Computer Society, Washington, DC, USA (2008), http://dx.doi.org/10.1109/AVSS.2008.21
6. Orosz, G., Novák, A.: Purepos 2.0: a hybrid tool for morphological disambiguation. In: RANLP'13. pp. 539–545 (2013)
7. Papazoglou, M.P., Heuvel, W.J.: Service oriented architectures: Approaches, technologies and research issues. The VLDB Journal 16(3), 389–415 (Jul 2007), http://dx.doi.org/10.1007/s00778-007-0044-3
8. Parker, L.E.: Task-oriented multi-robot learning in behavior-based systems. In: Intelligent Robots and Systems '96, IROS 96, Proceedings of the 1996 IEEE/RSJ International Conference on. vol. 3, pp. 1478–1487 vol.3 (Nov 1996)
9. Prószéky, G., Indig, B.: Natural parsing: a psycholinguistically motivated computational language processing model. In: 4th International Conference on the Theory and Practice of Natural Computing. Mieres, Spain (2015)
10. Siklósi, B., Novák, A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. XII. Magyar Számítógépes Nyelvészeti Konferencia pp. 3–14 (2016)
11. Specia, L., Paetzold, G., Scarton, C.: Multi-level translation quality prediction with quest++. In: ACL-IJCNLP 2015 System Demonstrations. pp. 115–120. Beijing, China (2015), http://www.aclweb.org/anthology/P15-4020
12. Specia, L., Shah, K., de Souza, J.G., Cohn, T.: Quest - a translation quality estimation framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 79–84. Sofia, Bulgaria (2013), http://www.aclweb.org/anthology/P13-4014
13. Stolcke, A.: Srilm - an extensible language modeling toolkit. pp. 901–904 (2002)
14. Yang, Z.G., Laki, J.L., Siklósi, B.: HuQ: An english-hungarian corpus for quality estimation. In: Proceedings of the LREC 2016 Workshop - Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem
15. Yang, Z.G., Laki, L.J., Siklósi, B.: Quality estimation for english-hungarian with optimized semantic features. In: Computational Linguistics and Intelligent Text Processing. Konya, Turkey (2016)