Gépi szövegértés vektoros nyelvmodellekkel

Kutatási terv

Makrai Márton

Bevezetés

A nyelvtechnológiában az utóbbi évtizedben egyre uralkodóbbá váltak a statisztikai módszerek. Ezek egyike az n-gram modell, ami sok tekintetben nagyon hasznosnak bizonyult, de egyben azzal a korláttal is bír, hogy nem tudja kihasználni a különféle szóalakok közötti hasonlóságot. Ezért ma egyre népszerűbbek a *vektoros nyelvmodellek* (Bengio et al., 2003), amelyek egy néhányszáz-dimenziós térben reprezentálják a szavakat, a hasonló szavakat egymáshoz közeli vektorokkal. A modellek gépi tanulással készülnek (gyakran neurális hálókkal, ma már több milliárd szavas korpuszokból is néhány óra alatt (Mikolov et al., 2013a)), így a vektorok koordinátáinak nem feltétlenül van szemléletes jelentése, mindenesetre a modell a szavak több tulajdonsága szerinti általánosításokat tanul meg. A modellek a nyelvtechnológia számos területén *state of the art* vagy ahhoz közeli eredményt adnak, így a hagyományos módszerek komoly versenytársai, rengeteg publikációval a témában. Összefoglalónak lásd Bengio et al. (2014) nyelvtechnológiai szakaszát. Mindemellett a modellekben kódolt szabályszerűségek intuitív értelmezése igazi alapkutatási feladat.

A kutatás keretében a következő három területen szeretnék előrelépni a vektoros nyelvmodellek segítségével: szavak közti jelentésviszonyok gyűjtése (I. szakasz), kompozicionális szerkezetek detektálása és reprezentálása (II. szakasz, a vizsgált szerkezetek a képzős szavak és a többszavas kifejezések) valamint fordítás és többjelentésűség problémaköre (III. szakasz). A terv további szakaszaiban sorra veszem ezeket a területeket: bemutatom a feladatot, leírom a tapasztalataimat és a terveimet. A kutatás legtöbb pontjának gyakorlati hasznához hozzájárul, hogy az utóbbi évek angol tárgynyelvű kutatásait magyarra alkalmazza. A kutatáshoz használt szoftvert és a keletkező adatbázisokat minden esetben szabadon hozzáférhetővé fogom tenni.

I. Lexikai relációk automatikus gyűjtése

A szövegek információtartalmának a zöme a szavakban van, így mind a hagyományos lexikográfiában, mind a gépi szövegértésben nagy hangsúlyt fektetnek a szavak közötti jelentésviszonyokra (lexikai relációk), mint a rokonértelműség, ellentétes jelentés (antonímia) vagy az okság (*bánt* → *fáj*). Számos lexikai relációval a nyelvtechnológusok sokat foglalkoznak. A szinonímia például az egyik legelterjedtebb számítógépes lexikonnak, a WordNetnek (Miller, 1995) az egyik fő rendezőelve. Az egyre nagyobb számítógéppel feldolgozható korpuszok korában ezeket a relációkat is automatikusan igyekszünk kinyerni. A szinonimák keresése a vektoros nyelvmodellek kiértékelésének is alapfeladata (*legközelebbi szomszédok*). Az antonímia vektoros modellekben való megjelenését mi is vizsgáltuk (Makrai et al., 2013), és azóta mások is (Zweig, 2014). Az okság vektoros nyelvmodellekben való megjelenésével kapcsolatban is végeztünk kísérleteket (Makrai, 2014).

Arról, hogy milyen viszonyokat érdemes vizsgálni, egy gépi szövegértési projekt (Kornai et al., 2015) szótárának (`4lang`, Kornai and Makrai (2013)) elkészítése során szereztem tapasztalatot, ahol bizonyos relációk, pl. a már említett okság vagy a birtoklás (‘az autónak kereke van’) kulcsszerepet játszanak. A kutatásban először feltérképezem, hogy mely lexikai relációk a legígéretesebbek a gépi szövegértéshez. Ebben olyan elméleti nyelvészeti megfontolások is szerepet játszanak, hogy különbséget kell-e tenni aközött a link között, ami a lexikon egy tételét annak okával illetve céljával kapcsolja össze, vagy inkább érdemes összeönteni az okokat és a célokkal. Egy másik megkülönböztetés, melynek hasznosságát vizsgálni kell, a birtokviszonyokat oszthatja fel inherensekre és elidegeníthetőkre. A relációkészletet kialakítását természetesen a relációk vektoros modellekből való minél jobb minőségű automatikus gyűjtése követi.

II. Képzők és többszavas kifejezések

Mind a tág értelemben vett nyelvleírásban, mind a számítópépes nyelvészetben elkülönül a lexikon a szabályoktól: az előbbibe kerülnek azok a nyelvi jelek, amiknek a jelentése nem megjósolható, az utóbbiak pedig a lexikon tételeiből felépítik a mondatokat (az alakjukat és természetesen a jelentésüket is). A feladat tehát két lépésből áll: a különféle egységekről meg kell állapítani, hogy kompozicionálisiak-e, ha nem, akkor eltárolni őket, ha pedig koncepcionálisak, akkor gondoskodni a jelentésüket kiszámító mechanizmusról. A projektnek ebben a szakaszában ennek megfelelően a vektoros nyelvmodellek segítségével először elkülönítjük a kompozicionális és nem-kompozicionális szerkezeteket, majd azzal foglalkozunk, hogy a lexikai tételek vektoros reprezentációjából hogyan számítható a nagyobb egységeké. Két területet választottunk: a szavaknál nagyobb nem-kompozicionális egységeket (II-A szakasz), a szavakon belül pedig a képzőket, melyek különböző mértékben kompozicionálisak (II-B szakasz).

II-A. Többszavas kifejezések

Többszavas kifejezések (kollokációk) gyűjtésére is alkalmasak a vektoros nyelvmodellek. A korpuszalapú lexikográfiában a kollokációk listáját a kifejezést alkotó szavak együttes előfordulásaiból számolt statisztikai mérőszámok alapján állítják össze, mint a Dice-együtthatóból számolt „lexikográfus-barát” logDice (Rychlý, 2008) vagy a pontonkénti

kölcsönös információ. A további szűrés intuitív vizsgálattal szokott történni. A vektoros nyelvmodellek lehetőséget adnak arra, hogy a potenciális többszavas kifejezésekhez is vektort rendeljünk. Dinu et al. (2013) egy olyan eszközt adtak közre, amivel többszavas kifejezések jelentését (vektorát) többféleképpen lehet számítani az őket alkotó szavak jelentéséből. Ebben a paradigmában fogom vizsgálni a *magyar* többszavas kifejezéseket. A munka során kikísérletezem, hogy a magyar korpuszokon mely kompozíciós függvények működnek a legjobban. Ez a kutatás tanulságokkal szolgál a kollokációkra vonatkozó nyelvészeti intuíció és a gépi tanulás terén is, és lexikográfiailag is hasznos.

A kollokációkra vonatkozó kutatás speciális esete az *igei szerkezetek*. Sass (2015) a saját adatán azt figyeli meg, hogy egy ige (pl. *száll*) kollokációi között szemléletesen úgy tudjuk elkülöníteni a kompozicionális szerkezeteket (*vonatra száll*) a komplex igéktől (*partra száll*), hogy az előbbiek argumentumai szemantikailag koherens osztályt alkotnak (a *vonat* helyett állhat *villamos*), az utóbbiak argumentumai (*part*) pedig kiugranak (*outlier*). Mivel a vektoros nyelvmodellekben kifejeződik a szavak szemantikai távolsága, a kompozicionális szerkezetek argumentumai (a *vonat* és a *villamos*) a vektortérben is közel vannak egymáshoz, a komplex igét alkotó argumentumok (*part*) pedig távolabb ezektől. A gyakori argumentumok vektorok alapján való klaszterezésével illetve a kiugrók megállapításával szeretném elkülöníteni az intuitíve különböző szerkezeteket. (Korábban is foglalkoztam predikátumok és bővítmények viszonyával a már említett `4lang` mélyesetrendszerének (Makrai, 2014) kidolgozása során.)

II-B. Képzők

A vektoros nyelvmodellek leggyakrabban szóalakokkal dolgoznak. A gépi tanulás velejárója, hogy a gyakoribb szavak reprezentációja pontosabb. Lazaridou et al. (2013) ezért a *képzett szavakhoz* (amelyek értelemszerűen sokkal ritkábbak, mint a relatív tövük) kompozicionálisan rendelnek vektort a tő és a képző reprezentációjából. Ha a képzés valóban kompozicionális, a felbontásból kapott reprezentáció jobb, mint a képzett szó előfordulásaiból tanult vektor.

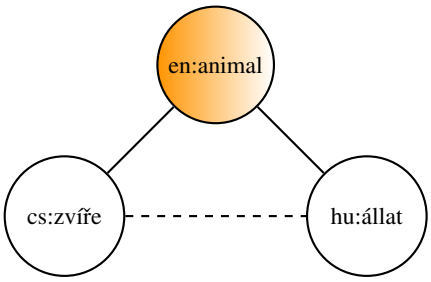
Szeretnék hasonló kutatást végezni magyarra, melynek *elméleti* hozadéka, hogy korpusz segítségével tudjuk elkülöníteni azt, hogy az egyes képzett szavakban a képző kompozicionális-e vagy csak történetileg van jelen. A várható *gyakorlati* haszon az angolhoz hasonlóan egy jobb minőségű vektoros nyelvmodell. Lazaridouék megjegyzik, hogy a fosztóképzők kezelését (*-less, un-*) nehezebbnek találták, ezért, ha az idő megengedi, szeretném az angol fosztóképzőkkel kapcsolatos nehézséget újragondolni az I. szakaszban leírt, az antonímiára vonatkozó kutatások eredményeinek a fényében.

III. Nyelvközi alkalmazások és egyértelműsítés

Az utolsó részfeladat a többnyelvűség világába vezet.

III-A. Többértelműség

A gépi szótárgenerálás egyik fő problémája az azonosalakúság (*homonímia*). Az angol *what* és *we* szavak magyar fordítása egyaránt *mi*, noha a két jelentésnek egyik nyelvben sincs köze egymáshoz. Kapcsolódó probléma a poliszémia, amikor az egyik nyelv különbséget tesz ott, ahol a másik nem, pl. a magyar *ablak* a németben *Fenster*, ha az épületen kívülre nyílik, és *Schalter*, ha hivatalnok ül mögötte.



1. ábra. Háromszögelés

Ezek a jelenségek (főleg a homonímia) sokat rontanak az automatikusan generált szótárakon. Az egyik nemkívánatos jelenség a gépi szótárgenerálás bevett eszközéhez, az úgynevezett háromszögeléshez kötődik. A háromszögelés azt jelenti, hogy abból, hogy a cseh *zvíře* angol fordítása *animal*, az *animal* magyar fordítása pedig *breaslă*, a 1 ábrán látható módon arra lehet következtetni, hogy a *zvíře* magyarul *állat*. A homonímia viszont hamis háromszögeket ad (német *was* – magyar *mi* – angol *we*).

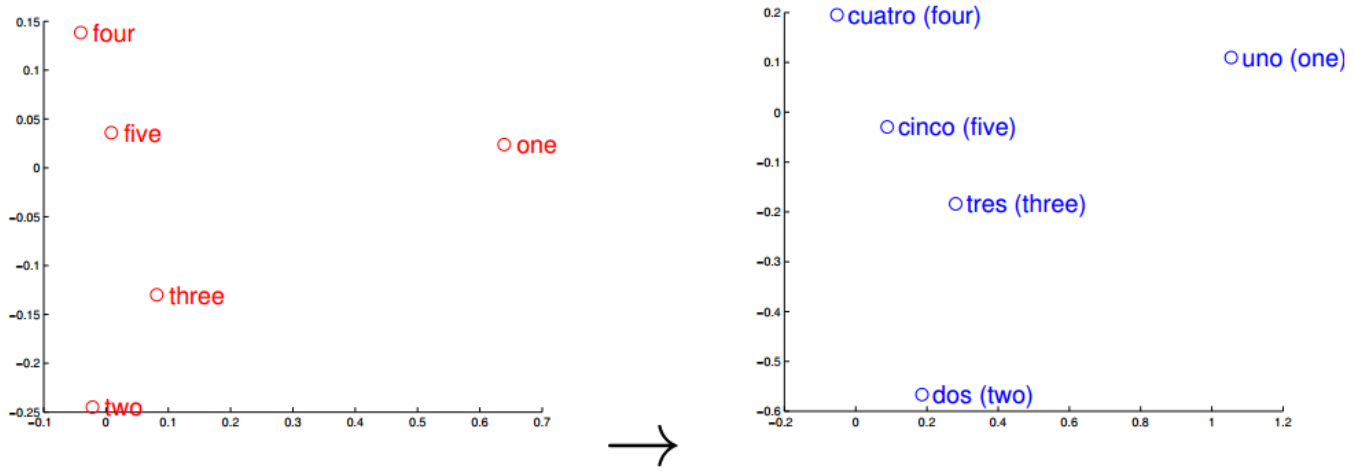
III-B. A probléma vektoros szemmel

A különböző nyelvek vektoros modelljei között olyan hasonlóságok vannak, amik gépi szótárgenerálásra használhatók (Mikolov et al., 2013b) (lásd a 2 ábrát. Mikolovék munkája a felügyelt gépi tanulás körébe tartozik: a kétnyelvű szótár generálásához adott egy-egy pár milliárd szavas egynyelvű korpusz a két nyelven, valamint egy néhány ezres kétnyelvű (*mag*) szótár. Az előbbiek segítségével elkészítjük a két nyelv vektoros modelljét, a magszótár segítségével pedig betanítunk egy lineáris leképezést a két vektortér között, ami a forrásnyelvi szavak vektorát a fordításuk vektorához közeli pontba viszi. A leképezés használható a teljes szókincs (valamilyen minőségű) lefordítására és meglevő fordítások pontozására is. A módszert magyarra is alkalmaztuk (Makrai, 2015).

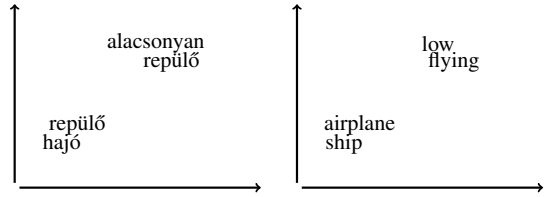
A homonímia (és a poliszémia) a vektoros szótárgenerálás minőségét is rontja, mert a nyelvmodellek a legegyszerűbb esetben a szóalakokhoz rendelnek egy-egy vektort, így többjelentésű szavak esetén a vektorban a különféle jelentések keverednek.

III-C. A megoldás

Az előbb említett problémákra két megoldást is szeretnék megvizsgálni. Az egyik azon alapszik, hogy míg a hamis háromszögeket a háromszög közepét képező nyelvben levő homonímia okozza, a vektoros módszer csak a forrás- és a célnyelv többértelműségeire érzékeny, így a kettő kompenzálhatja egymást. Olyan szótárakból szeretném kiszűrni a hamis háromszögeket Mikolov et al. (2013b) módszerével, amiket a Wiktionary nevű közösségi szerkesztésű soknyelvű



2. ábra. A különböző nyelvek vektoros modelljei hasonlóak. Ez teszi lehetővé Mikolovék módszerének alkalmazását.



3. ábra. Szójelentések fordítása Mikolovék módszerével. A *repülő* szóalak igenév és (a nyelvtörténeti körülményeket nem tekintve) egymorfémás főnév is lehet. A két jelentése hasonló szavak azonosítják be, mint a főnévi értelemben szinonim *hajó* illetve az igenévvel kollokáló *alacsonyan*

szótárból gyűjtünk (Ács et al., 2013). Magszótárként közvetlen, felhasználók által megadott fordításokat fogok használni, és az ezekkel tanított lineáris leképezés segítségével vizsgálom a háromszögeléssel kapott fordításokat. Azt várom, hogy a homonímiából adódó hamis párok így kiszűrhetők.

A másik lehetőség a homonímia és a poliszémia kezelésére a vektoros nyelvmodelleken belüli. Reisinger and Mooney (2010) és Huang et al. (2012) olyan vektoros nyelvmodellt tanítanak be, ami a szavak különböző jelentéseit külön-külön vektorral (a pszichológiai fogalommodellezés szavával élve prototípussal) reprezentálja. A paradigmát továbbfejlesztette és hatékony nyílt forráskódú implementációval látta el Bartunov et al. (2015). Az általam tervezett kutatás újdonsága, hogy a többértelműségre érzékeny modelleket fordításban alkalmazom. A különböző jelentéseket hasonló szavakkal azonosítom be, és mindegyikhez fordítást adok Mikolovék módszerével, lásd a 3 ábrát.

III-D. Elméleti hozadék: a jelentéskészlet kritikája

A poliszémia kezelésének kutatásából elméleti nyelvészeti tanulságokat is le szeretnék vonni. A magyar-német szótárakban fel szokták tüntetni, hogy az *ablak* szónak két „jelentése” van: *Fenster* és *Schalter*. A nyelvtanuló számára ez hasznos, a szemantika strukturalista felfogása szerint azonban a szavak jelentését oppozíciók definiálják, és ebből a szempontból kérdéses, hogy a magyar *ablak* szó valóban többértelmű-e. Ruhl (1989)-es programmatikus könyvében azt javasolja, hogy egy szóalaknak lehetőleg csak egy (absztrakt) jelentést tulajdonítsunk, a szó különböző előfordulásai közötti különbségeket pedig a kontextusból, a környező szavak (ugyancsak absztrakt) jelentéséből vezessük le. Az emberi beszélők ilyen téren nagyon kreatívak, a metaforikusnak tekintett kifejezések alkotása és megértése a nyelvi készség egyik legalapvetőbb része, és az ezeket lehetővé tevő jelentésreprezentáció a gépi szövegértés számára is kívánatos. A poliszémiát kvantitatíve szeretném vizsgálni egy a többértelműségre érzékeny modellben *Fenster–Schalter* típusú szópárokkal, a két szó vektora közötti távolság alapján.

Végül megjegyzem, hogy Mikolov et al. (2013c) megmutatták, hogy vektorok közötti különbség(vektor)nak is van fogalmi jelentése.

$$\mathbf{king} - \mathbf{queen} \approx \mathbf{man} - \mathbf{woman}$$

A *Fenster–Schalter* típusú pároknál kapott különbségvektorokat szeretném megvizsgálni ebből a szempontból is.

Hivatkozások

Judit Ács, Katalin Pajkossy, and András Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. ACL.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. *ArXiv preprint*, 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. 2014. URL http://arxiv.org/abs/1206.5538.

G Dinu, N Pham, and M Baroni. Dissect: Distributional semantics composition toolkit. In *Proceedings of the System Demonstrations of ACL*, 2013.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390524.2390645.

András Kornai and Márton Makrai. A 4lang fogalmi szótár. In Attila Tanács and Veronika Vincze, editors, *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 62–70, 2013.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S15-1019.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526, 2013.

Márton Makrai. Deep cases in the `4lang` concept lexicon. In Attila Tanács, Viktor Varga, and Veronika Vincze, editors, *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014)*, pages 50–57 (in Hungarian), 387 (English abstract), 2014. ISBN 978-963-306-246-3.

Márton Makrai. Comparison of distributed language models on medium-resourced languages. In Attila Tanács, Viktor Varga, and Veronika Vincze, editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, 2015. ISBN 978-963-306-359-0.

Márton Makrai, Dávid Márk Nemeskey, and András Kornai. Applicative structure in vector space models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 59–63, Sofia, Bulgaria, August 2013. ACL. URL http://www.aclweb.org/anthology/W13-3207.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *Proc. ICLR 2013*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya Sutskeve. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168, 2013b.

Tomas Mikolov, Wen-tau Yih, and Zweig Geoffrey. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT-2013*, pages 746–751, 2013c.

George A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.

C. Ruhl. *On monosemy: a study in lingusitic semantics*. State University of New York Press, 1989.

Pavel Rychlý. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, 2008.

Bálint Sass. 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet. In Tanács Attila, Varga Viktor, and Vincze Veronika, editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, 2015. ISBN 978-963-306-359-0.

Geoffrey Zweig. Explicit representation of antonymy in language modeling. Technical report, Microsoft Research, 2014.