Elosztott nyelvmodellek összehasonlítása közepes nyelveken

Makrai Márton

MSZNY 2015

Áttekintés

A feladat

- szótárgenerálás, EFNILEX
  - lexikográfiához a gépi fordítás eszközeivel
  - kisebb európai nyelvekre ($\leq$ magyar)
- ötlet (Mikolov et al., 2013b)
  - két egynyevlű korpusz + néhányezres seedszótár
  - a két nyelv vektoros modellje közötti lineáris leképezés

Vektoros nyelvmodell

- = *continuous vector space model, distributed language model*
- minden szóalakhoz egy 200–800 dimenziós sűrű valós vektor
- hasonló szavak közel (cos)
- relációs hasonlóság

$$\textbf{lerövidíti} - \textbf{lerövidít} \approx \textbf{mondja} - \textbf{mond}$$

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

- alkalmazás: nyelvmodellezés, címkézési feladatok, képi vektorokkal összekötve, gépi fordítás, parafrázisfelismerés, szójelentés-egyértelműsítés, véleményelemzés

Neurális hálók és együtt-előfordulások

- nyílt forráskódú, nagyon hatékony eszközök
- word2vec (Mikolov et al., 2013a)
  - eredetileg C-ben
  - gensim, pythonos optimalizálás (Řehůřek and Sojka, 2010)
  - LBLword2vec (a word2vec levlistáján)
- GolVe (Pennington et al., 2014), kapcsolat a mátrixfaktorizációval

Fordítási mátrix

- tanulás a leggyakoribb 5K szón

$$\min_W \sum_{i=1}^{5000} ||Ws_i - t_i||^2$$

- teszt a következő 1K szón

$$\arg\max_t \cos(Ws_i, t)$$

- két különböző távolság
- a tesztnél a hiányzó fordítású szavakat kihagyjuk

Egynyelvű előfeladat

-
  - közös reprezentáció (*representation sharing*)
  - gyorsabb teszt
- analógiás kérdések (Mikolov et al., 2013d)
  - a *man* úgy aránylik a *woman*-hez, mint a *king* mihez?

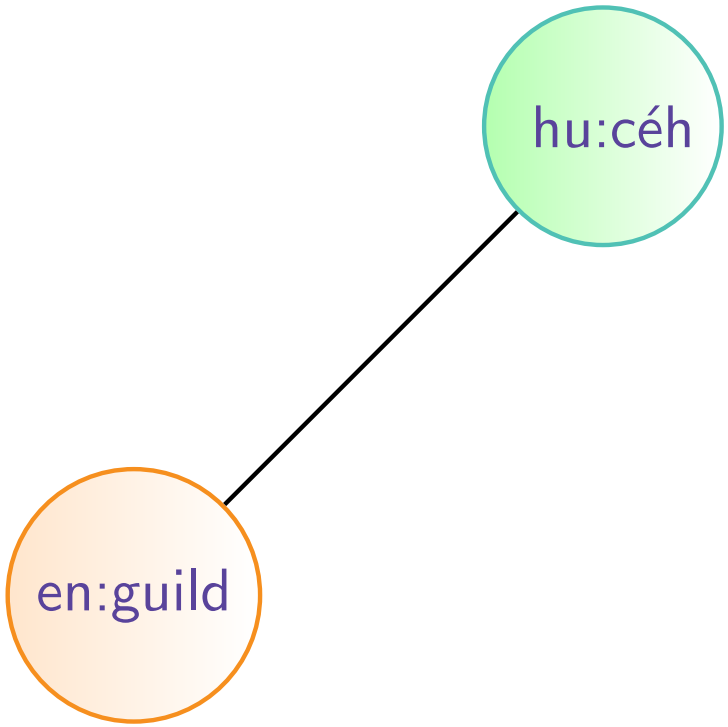
$$\mathbf{man} - \mathbf{woman} \approx \mathbf{king} - x$$

$$x \approx_{\cos} \mathbf{king} - \mathbf{man} + \mathbf{woman}$$

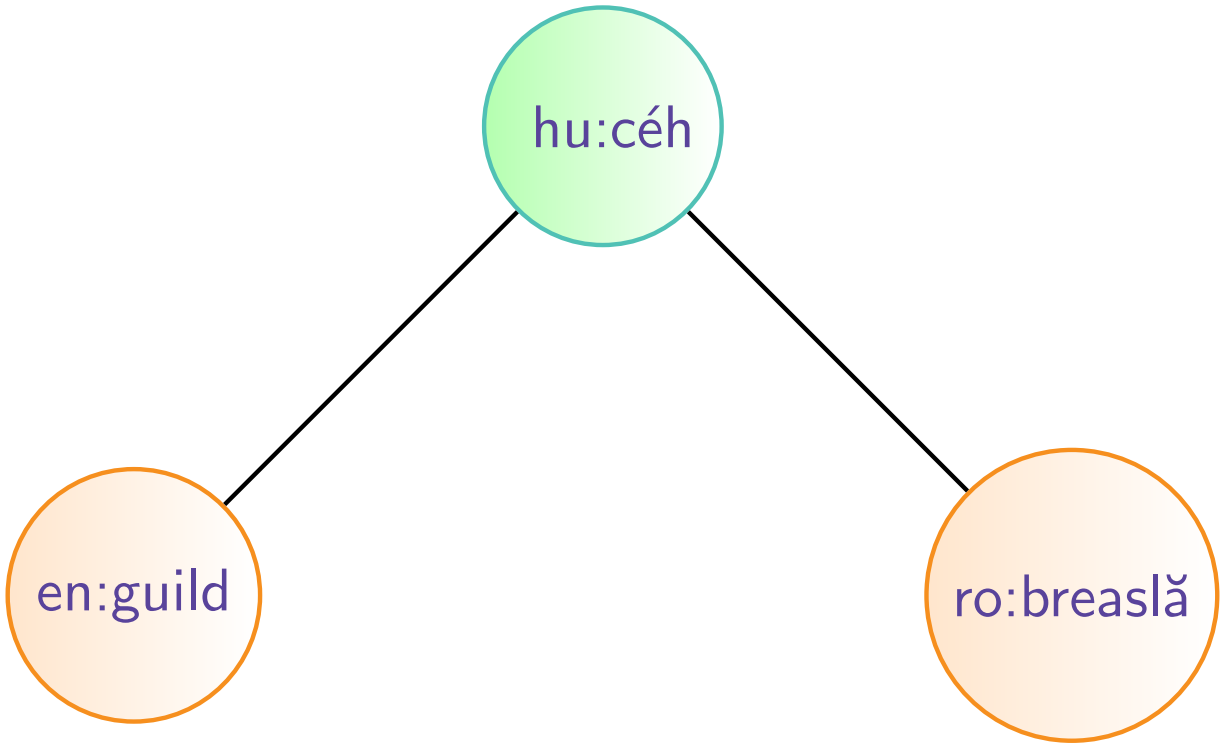
Áttekintés

- korpusz vagy készen vett nyelvmodell
  - angol: modellek az eszközök honlapjáról
  - magyar: MNSZ2 (Oravecz et al., 2014), Webkorpusz (Halácsy et al., 2004)
  - szlovén (Ljubešić and Erjavec, 2011) (hr 1.9B, sl 1.2B, sr 0.9B, bs 0.4B, (Ljubešić and Klubička, 2014))
  - litván (Zséder et al., 2012)
- seedszótár
  - efnilex12 (Héja and Takács, 2012)
  - opus (Tiedemann, 2012)
  - Wiktionary $\pm$ háromszögelés (Ács et al., 2013)

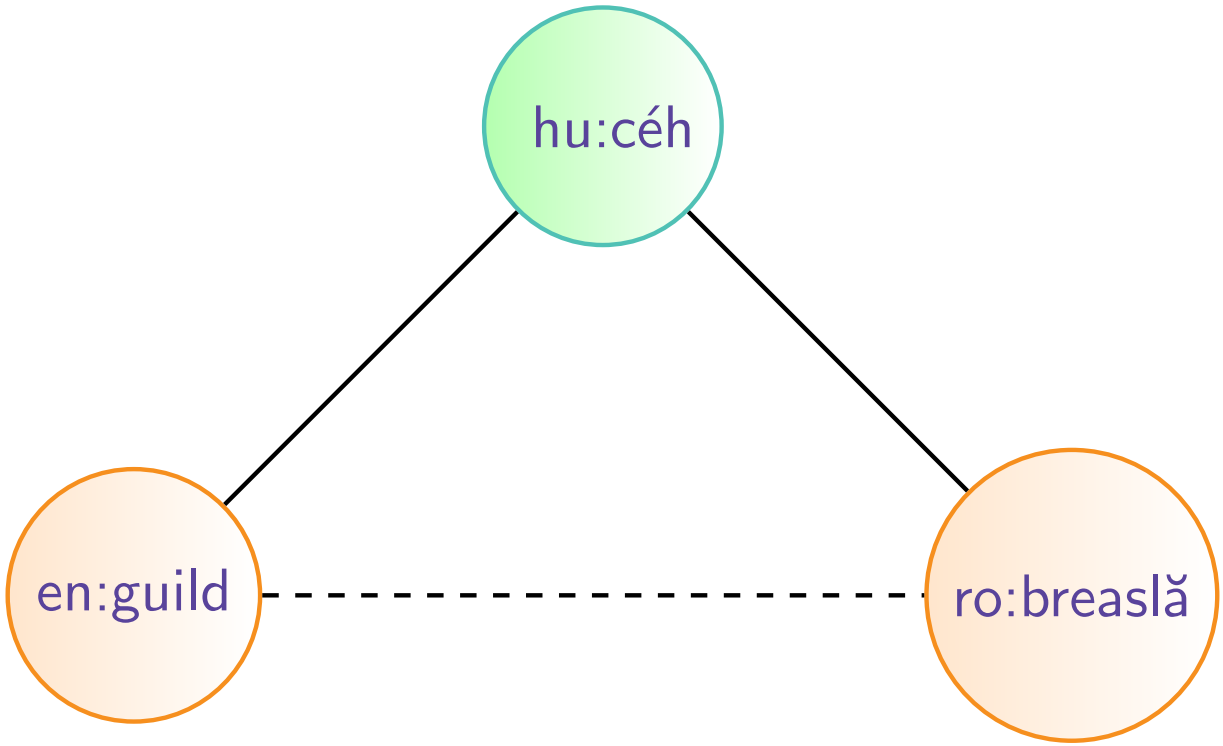
Wiktionary ± háromszögelés



Wiktionary ± háromszögelés



Wiktionary ± háromszögelés



Áttekintés

Morfológiai szópárok

angol		magyar	
decrease	decreases	lesznek	lesz
describe	describes	állnak	áll
eat	eats	tudnak	tud
enhance	enhances	kapnak	kap
estimate	estimates	lehetnek	lehet
find	finds	nincsenek	nincs
generate	generates	kerülnek	kerül

a leggyakoribb példákat vettük a Webkorpusz szerint (Halácsy et al., 2004)

Szemantikai szópárok

angol		magyar	
Athens	Greece	Budapest	Magyarország
Baghdad	Iraq	Moszkva	Oroszország
Bangkok	Thailand	London	Nagy-Britannia
Beijing	China	Berlin	Németország
Berlin	Germany	Pozsony	Szlovákia
Bern	Switzerland	Helsinki	Finnország
Cairo	Egypt	Bukarest	Románia

Analógiás kérdések

angol				magyar			
Athens	Greece	Baghdad	Iraq	Budapest	Magyarország	Moszkva	Oroszország
Athens	Greece	Bangkok	Thailand	Budapest	Magyarország	London	Nagy-Britannia
Athens	Greece	Beijing	China	Budapest	Magyarország	Berlin	Németország
Athens	Greece	Berlin	Germany	Budapest	Magyarország	Pozsony	Szlovákia
Athens	Greece	Bern	Switzerland	Budapest	Magyarország	Helsinki	Finnország
Athens	Greece	Cairo	Egypt	Budapest	Magyarország	Bukarest	Románia

Feladattípus	angol		magyar
	# kérdés	# pár	# kérdés
gram1-adjective-to-adverb	32	992	40
gram2-opposite	812	29	30
gram3-comparative	37	1332	40
gram4-superlative	34	1122	40
gram5-present-participle	33	1056	40
gram6-nationality-adjective	41	1599	41
gram7-past-tense	40	1560	40
gram8-plural-noun	37	1332	40
gram9-plural-verb	30	870	40
capital-common-countries	23	506	20
capital-world	116	4524	166
city-in-state	2467	68	
county-ceter			19
county-district-center			175
currency	30	866	30
family	23	506	20

Áttekintés

Analógiás kérdések

- korpusz: MNSZ2 $\oplus$ Webkorpusz
- a többi paraméter Mikolov et al. (2013c) szerint (300-dimenziós) word2vec-sgram

Analógiás kérdések

- korpusz: MNSZ2 $\oplus$ Webkorpusz
- a többi paraméter Mikolov et al. (2013c) szerint (300-dimenziós) word2vec-sgram

		morf		szemant	
en	$n = 5$	61		58	
	$n = 15$	61		61	
	HS	52		59	
hu	$n = 5$	63.0	3419/5430	38.5	269/699
	$n = 15$	61.9	3359/5430	39.2	274/699
	HS	48.9	2653/5430	22.5	157/699

Szófordítás

	prec@1	prec@5
en $\rightarrow$ sp	33	51
sp $\rightarrow$ en	35	52
en $\rightarrow$ cz	27	47
cz $\rightarrow$ en	23	42
en $\rightarrow$ vn	10	30
vn $\rightarrow$ en	24	40
glove-sl $\rightarrow$ en rs	44.80	63.40
word2vec-sl $\rightarrow$ en $m = 100$ rs	41.70	60.40
word2vec-hu $\rightarrow$ en $m = 50$ rst	32.80	54.70
word2vec-lt $\rightarrow$ en $m = 100$ rt	21.20	36.50

Hasonlósági mérték

- A pontosság és fedés csereviszonya

cos >	szókincs	gold	prec@1	prec@5
0.75	1440	100	75.0%	86.0%
0.7	3803	301	68.4%	84.4%
0.65	6931	516	60.9%	79.5%
0.6	9967	711	54.7%	74.1%
0.55	12008	884	48.9%	68.7%
0.5	12949	958	46.6%	65.6%
0.45	13300	981	45.7%	64.3%
0.4	13451	988	45.3%	64.0%
0.35	13511	993	45.1%	63.8%
0.32	13520	994	45.1%	63.8%

- adott szótárban levő hibásan fordított vagy többértelmű szavak felderítésére

tekintenek	0.6618	concerned	assume	themselves	merely	i
oldalát	0.6618	side	sides	bottom	inside	r
nemrégiben	0.6618	recently	announced	recent	earlier	r
kaphat	0.6618	get	give	wants	giving	a
hivatalba	0.6618	election	elections	elected	council	p
fordításban	0.6618	translated	translations	translation	writings	w
fekvésű	0.6618	situated	spacious	south	overlooking	e
érve	0.6618	pushed	down	drove	behind	a
emlékeztetni	0.6618	perhaps	mention	forget	however	r
előadást	0.6618	lecture	presentation	performances	audience	l
csomagok	0.6618	packages	customers	package	mail	a
washingtoni	0.6617	President	government	Washington	U.S.	C
tiszteletét	0.6617	dignity	reverence	tradition	honor	f
Számítástechnikai	0.6617	Computer	Systems	Engineering	Technical	T

- Tervek
  - tövezett korpusszal
  - többszavas kifejezések
  - egyértelműsítés
- `http://corpus.nytud.hu/efnilex-vect/`

Hivatkozások I

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, 2013. Association for Computational Linguistics.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 203–210. ELRA, 2004.

Enikő Héja and Dávid Takács. An online dictionary browser for automatically generated bilingual dictionaries. In *Proceedings of EURALEX2012*, pages 468–477, 2012.

Nikola Ljubešić and Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer, 2011.

Nikola Ljubešić and Filip Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Hivatkozások II

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013c. URL `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositiona` `pdf`.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia, 2013d. Association for Computational Linguistics.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. The Hungarian Gigaword Corpus. In *Proceedings of LREC 2014*, 2014.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

Hivatkozások III

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. URL `http://is.muni.cz/publication/884893/en`.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, editor, *LREC*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Attila Zséder, Gábor Recski, Dániel Varga, and András Kornai. Rapid creation of large-scale corpora and frequency dictionaries. In *Proceedings to LREC 2012*, pages 1462–1465, 2012.