Okság vektoros nyelvmodellben

Makrai Márton

MTA Nyelvtudományi Intézet
témavezető: Kornai András

Tavaszi Szél Konferencia 2014

Áttekintés

Gépi szövegértés

Gépi szövegértés

Az IBM *Watson*-ja megnyeri a *Jeopardy*-t (2011)

Nyelvtechnológia

Nyelvtechnológia

- felhasználói szinten
  - webes keresés
  - gépi fordítás
  - optikai karakterfelismerés
  - beszédfelismerés és -generálás
  - információkinyerés folyó szövegből
  - érzelmi elemzés
  - szöveg összegzése, egyszerűsítése
- és ami mögötte van
  - toldalékok megállapítása, szótövezés
  - szófaji címkézés
  - szavak egyértelműsítése
  - tulajdonnév-felismerés
  - mondattani elemzés
  - nyelvmodellezés

Áttekintés

n-gramm modellek

n-gramm modellek

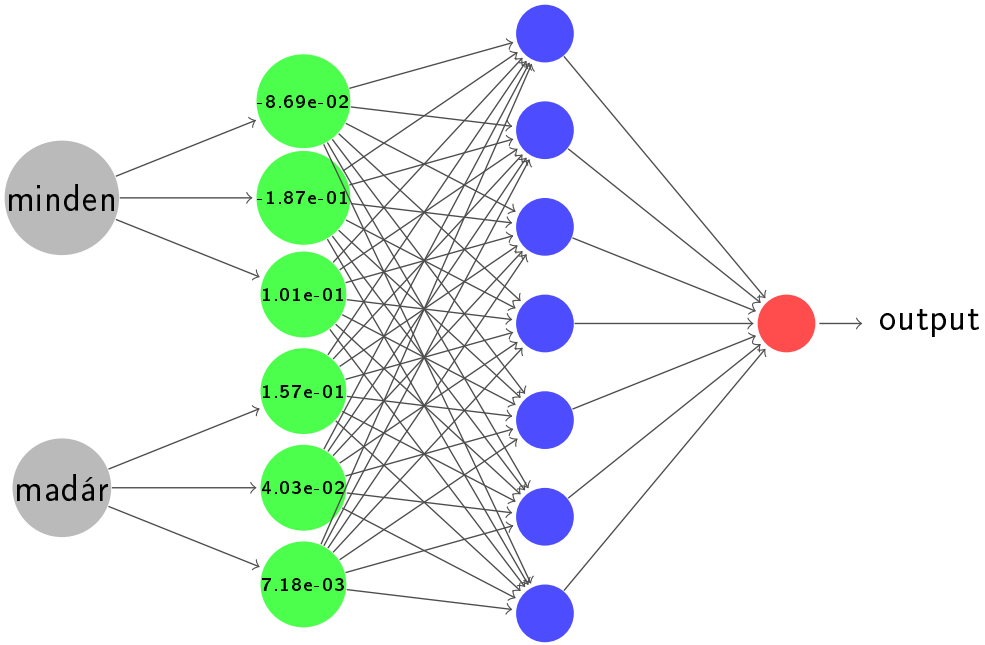
- sztring pl.
  - karakter → szó
  - szó → mondat
  - mondat → dialógus vagy bekezdés
- sztringek (véges hosszú sorozatok) valószínűségét modellezi
- a valószínűségeket relatív gyakoriságokkal közelítjük
- végtelen sok sztringet kell modellezni véges minta alapján: általánosítani kell
- *n*-gramm modell, Markov-tulajdonság

$$P(\texttt{\^{}\ minden\ madár\ társat\ választ\ \$}) =$$

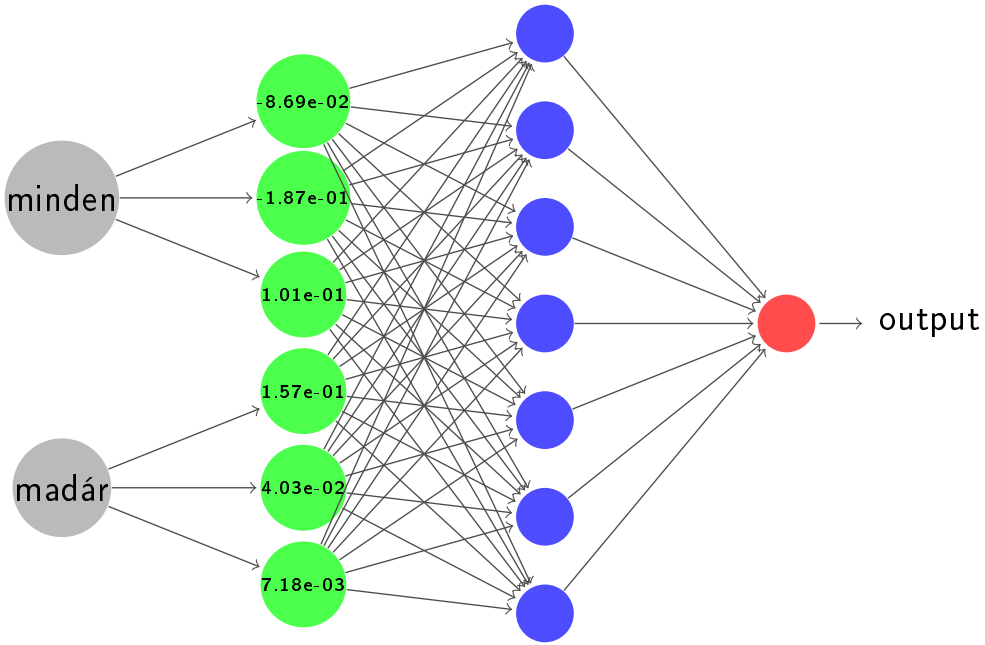
$$P(\ \texttt{minden}\ |\ \texttt{\^{}}\ ) \cdot P(\ \texttt{madár}\ |\ \texttt{minden}\ ) \cdot P(\ \texttt{társat}\ |\ \texttt{madár}\ ) \cdot$$

$$\cdot P(\ \texttt{választ}\ |\ \texttt{társat}\ ) \cdot P(\ \texttt{\$}\ |\ \texttt{választ}\ )$$

Vektoros nyelvmodellek



Vektoros nyelvmodellek



- Vektoros nyelvmodellek számítása:
  - legtöbbször a szavak különféle kontextusokban való előfordulásának gyakoriságát tartalmazó mátrixból számítják
  - máskor neurális hálókkal tanulják
  - mi olyan nyelvmodelleket is nézünk, amelyek a 4lang fogalmi szótárból készültek

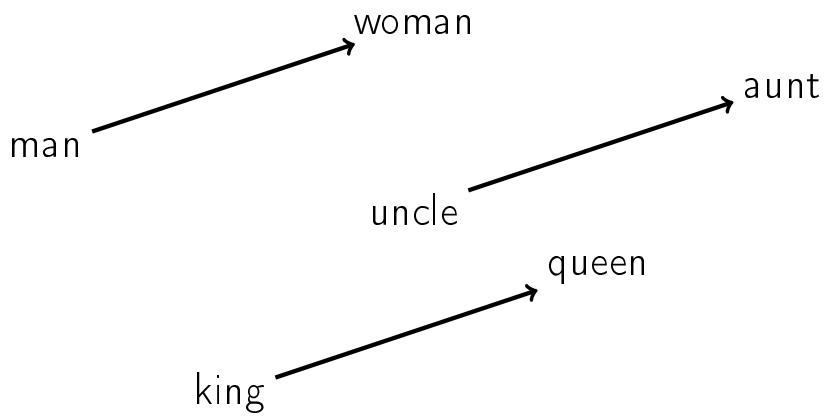
A 41ang fogalmi szótár (Kornai and Makrai, 2013)

A 41ang fogalmi szótár (Kornai and Makrai, 2013)

- a tételek az egyes nyelvek szavainál absztraktabb fogalmak
  - a szótárakban megszokottnál jóval kevesebb esetben bontunk szét egy szót több jelentésre, pl. ablak (a szabadba nyílik, vagy jegykiadó)
  - többnyelvű tételek és nyelvfüggetlen(nek szánt) definíciók (több nagyon különböző nyelvet is figyelembe veszünk)
  - szófajbeli különbségek nem befolyásolják a jelentést, az *engedély* és a *megenged* jelentése ugyanaz
- a tételek jelentése definiálva van egy formális nyelven
- szabadon letölthető

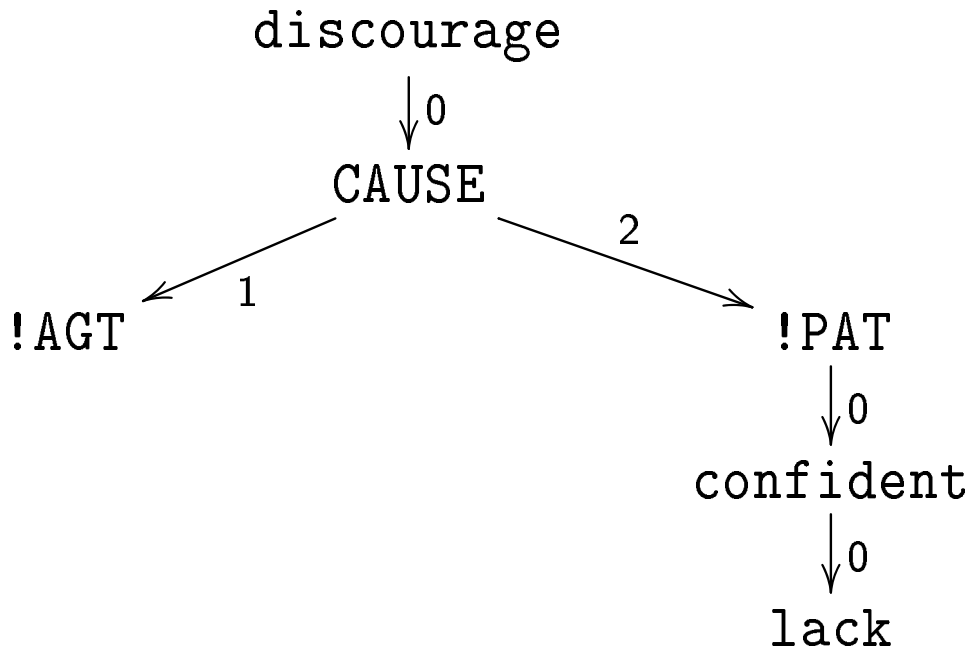
Áttekintés

A vektoros eltolás módszere (*vector offset method*, (Mikolov et al., 2013))



$$\mathbf{v}\left(\text{king}\right) + \mathbf{v}\left(\text{female}\right) = \mathbf{v}\left(\text{queen}\right)$$

Okság a jelentésreprezentációban



Okok és okozatok a szókincsben

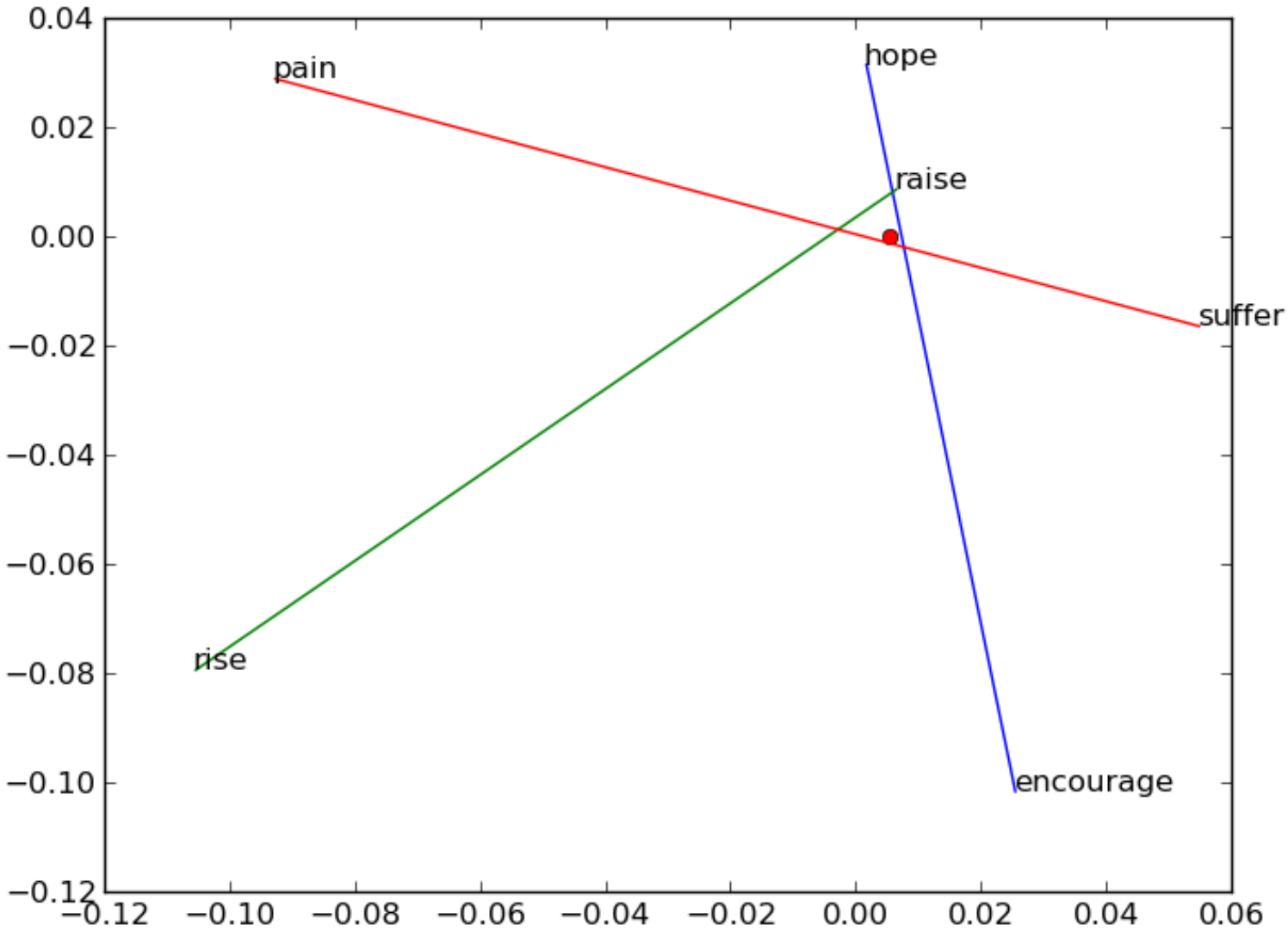
- az ok–okozat párokat a WordNet-ből (Miller, 1995) vettem
  - *synset*ek
  - glossza (szöveges definíció)
  - példamondat
  - szemantikai viszonyok, pl. is-a, instance-of, antonym, part-of, cause

Okok és okozatok a szókincsben

- az ok–okozat párokat a WordNet-ből (Miller, 1995) vettem
  - *synset*ek
  - glossza (szöveges definíció)
  - példamondat
  - szemantikai viszonyok, pl. is-a, instance-of, antonym, part-of, cause

give	have
show	see
encourage	hope
feed	eat
kill	die
raise	rise
⋮	⋮

Áttekintés


pain
hope
raise
suffer
rise
encourage
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
−0.12
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06

- Statisztikai tesztek
  - kiszámítjuk a párokra illesztető egyenesekhez legközelebbi $\mathbf{c}$ pontot (centrum) Han and Bancroft (2010)
  - a pároknak $\mathbf{c}$-től való X távolságát nézzük
  - az igazi párok (okok és megfelelő okozatok) vs véletlen párok
  - $\mathbb{E}X_i \overset{?}{<} \mathbb{E}X_v$
  - a különbség szignifikanciája: kétmintás t-próba
- eredmények
  - centrális:
    - SENNA, Collobert et al. (2011), $p < 0.001$
    - a 4lang nyelvmodellek némelyikénél, $p < 0.05$
  - mások nem centrálisak (Turian et al. (2010); Huang et al. (2012), HLBL Mnih and Hinton (2009), English Polyglot Al-Rfou' et al. (2013))

Nyitott kérdések

Nyitott kérdések

- összefüggés a nyelvmodell számítása és a centralitás között
- bonyolultabb geometriai modell gépi tanulása
- az okozat helye az ok és a középpont egyenesén
  - $\mathbf{v}\,(\text{ok})$ és $\mathbf{c}$ közötti pontok az egyenesen

$$\mathbf{v}\,(\text{okozat}) = \lambda \mathbf{v}\,(\text{ok}) + (1-\lambda)\mathbf{c}$$

  - általánosabban keressük:

$$\mathbf{v}\,(\text{okozat}) \approx \lambda_{\text{ok}} \mathbf{v}\,(\text{ok}) + \lambda_c \mathbf{c}$$

  - az egyes pároknál optimális együtthatók átlagát használva közelítettem az $\mathbf{v}\,(\text{okozat})$-ot az egyes $\mathbf{v}\,(\text{ok})$-oknál.
  - probléma: az okozat mindig nagyon közel van $\mathbf{c}$-hez

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-3520`.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 2011.

Lejia Han and John C. Bancroft. Nearest approaches to multiple lines in n-dimensional space. In *CREWES Research Report*, volume 22. University of Calgary, 2010.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 873–882, Jeju Island, Korea, 2012. Association for Computational Linguistics.

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

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, 2013. Association for Computational Linguistics.

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

Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21: 1081–1088, 2009.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, 2010. Association for Computational Linguistics.