TEXT MINING 2

Objavovanie znalostí v textoch

Peter Bednár

Počítačová lingvistika

Počítačová lingvistika

- Interdisciplinárna oblasť medzi lingvistikou, umelou inteligenciou a informatikou
- Zaoberá sa návrhom počítačových algoritmov pre spracovanie prirodzeného jazyka (hovorenej reči alebo textového prejavu)
- Vytvára formálne modely pre jednotlivé úrovne prirodzeného jazyka

Úrovne jazyka (1)

- Spracovanie je rozdelené na čiastkové úlohy podľa jazykových úrovní:
  1. **Lexikálna úroveň**
  2. **Morfologická úroveň**
  3. **Syntaktická úroveň**
  4. **Sémantická úroveň**
  5. Pragmatická úroveň

Úrovne jazyka (2)

1. **Lexikálna úroveň**
   - Ako rozdeliť postupnosť znakov na základné jazykové jednotky - tokeny
2. **Morfologická úroveň**
   - Aké sú vlastnosti jednotlivých slov
3. **Syntaktická úroveň**
   - Aké sú vzťahy medzi slovami podľa ich poradia vo vete
4. **Sémantická úroveň**
   - Aká je jednoznačná reprezentácia významu
5. **Pragmatická úroveň**
   - Čo autor textu zamýšľa a akú akciu má počítač vykonať

Úlohy pri lingvistickej analýze

- Lexikálna analýza
  - Tokenizácia – rozdelenie na základné lexikálne jednotky
  - Segmentácia textu – rozdelenie na vety, vyjadrenia
- Morfologická analýza
  - Lematizácia – prevedenie tvarov na základný tvar slova
  - Morfologické značkovanie – priradenie slovného druhu (sloveso, podstatné meno, predložka, atď.) a morfologických kategórií (rod, pád, číslo, vid, atď.)
- Syntaktická analýza
  - Syntaktické parsovanie – rozdelenie vety na frázy, priradenie syntaktických rolí (podmet, prísudok, predmet, atď.)

Morfologické značkovanie

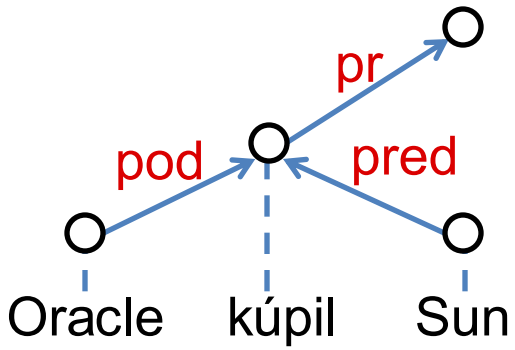
- Pri morfologickom značkovaní sa každému slovu priradí jeho slovný druh (podstatné meno, prídavné meno, sloveso, príslovka, predložka, zámeno, spojka, častica, citoslovce, atď.) a ďalšie morfologické kategórie (rod, pád, číslo, vid, životnosť, atď.)
- Na základe poradia slov a morfologických značiek sa potom identifikujú syntaktické vzťahy
- Na automatické značkovanie sa používajú metódy strojového učenia + ohraničenie morfologickým slovníkom – dosahujú presnosť 90-99% podľa zložitosti morfológie jazyka

Syntaktické parsovanie

- Pri syntaktickom parsovaní sa identifikujú gramatické vzťahy medzi slovami a vetné členy (podmet, prísudok, predmet, prívlastok, atď.)
- Na automatické parsovanie sa používajú metódy strojového učenia – dosahujú presnosť 90-96% podľa jazyka
- V súčasnosti je populárna reprezentácia pomocou závislostných stromov, ktorá je výhodná pre jazyky s voľným slovosledom (napr. pre slovenčinu)

Syntaktická reprezentácia - závislostné stromy

- Každému slovu vo vete zodpovedá jeden uzol + koreň stromu
- Orientované hrany medzi uzlami reprezentujú gramatické vzťahy a sú označené funkciou podradeného slova (počiatočný uzol) voči nadradenému slovu (koncový uzol hrany), napr.:



Syntaktická reprezentácia – príklad (1)

Slovné druhy

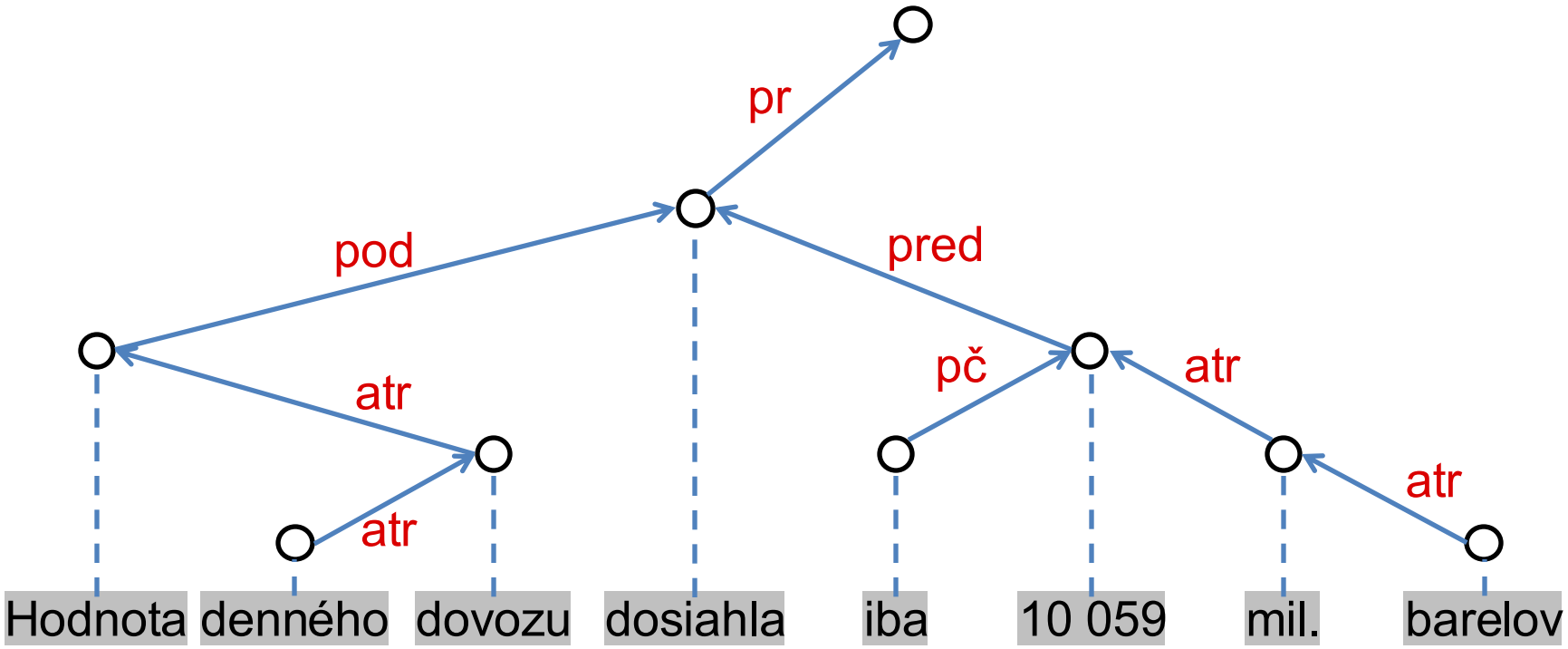
PM podstatné meno, PrM prídavné meno, S sloveso, ZČ zvýrazňovacia častica, Čsl číslovka

Morfologické kategórie

1p, 2p – pád, jč, mč – jednotné/množné číslo, žr, mr – ženský/mužský rod, nž – neživotné, dok – dokonavý slovesný vid, atď.

Hodnota	denného	dovozu	dosiahla	iba	10 059	mil.	barelov
PM 1p, jč, žr	PrM 2p, jč, mr	PM 2p, jč, mr, nž	S dok jč, žr	ZČ	Čsl	Čsl 2p, mč, mr	PM 2p, mč, mr, nž

Syntaktická reprezentácia – príklad (2)



Syntaktické funkcie

pr – prísudok, pod – podmet, pred – predmet, atr – atribút rozvitého člena, pč – pomocný člen

Sémantická reprezentácia

- Cieľom je vytvoriť jednoznačnú štruktúrovanú reprezentáciu významu textu
- Najpoužívanejšia je reprezentácia pomocou znalostných grafov
  - Uzly – jednoznačne definované pojmy/entity
  - Hrany – relácie medzi entitami
- Výsledný graf je možné reprezentovať ako množinu faktov v tvare subjekt-predikát-objekt – *tripletov*.

Znalostné grafy - príklad

- DBpedia
  - Popisuje všeobecné fakty vyextrahované z Wikipedie
  - 4.58 miliónov objektov klasifikovaných podľa typov (1 445 000 osôb, 735 000 miest, 241 000 organizácií, 411 000 umeleckých diel, atď.)
  - Celkovo 3 miliardy tripletov
  - http://wiki.dbpedia.org/

Konštrukcia znalostných grafov

- Extrahovanie pomenovaných entít
  - Extrahovanie názvov, ktoré sa v texte odkazujú na rôzne typy objektov, alebo hodnôt ich vlastností
- Rozlíšenie koreferencií
  - Rozlíšenie na ktoré objekty v texte odkazujú zámená a menné frázy
- Extrahovanie relácií
  - Extrahovanie vzťahov medzi entitami - výsledok je množina faktov v tvare subjekt-predikát-objekt

Extrahovanie entít

- Cieľom je nájsť v texte slová, alebo slovné spojenia, ktoré označujú pomenované entity
- Entity zodpovedajú uzlom znalostného grafu
- Napr. mestá, osoby, organizácie, udalosti ... produkty ... diagnózy, chemické zlúčeniny, DNA sekvencie, lieky ...

Rozlíšenie koreferencií (1)

- Cieľom je rozlíšiť na ktoré objekty v texte odkazujú zámená a menné frázy
- Zámená a nevyjadrené členy
  - Môžu sa vyskytovať pred, alebo po odkazovanom objekte
  - Môžu sa odkazovať aj na sloveso
  - Nejednoznačnosť, ktorá sa často nedá rozhodnúť na syntaktickej úrovni

Oracle kúpil firmu Sun. To[kúpa] bolo najdôležitejšie pre jej[Oracle? Sun?] rozvoj.

Rozlíšenie koreferencií (2)

- Menné frázy
  - Rozlíšenie vyžaduje vo všeobecnosti doménové znalosti

Karzai a Bush rokovali o Talibane. Americký prezident$_{[Bush]}$ sa stretol s afganským prezidentom$_{[Karzai]}$ na konferencii v Kábule.

Extrahovanie relácií

- Cieľom je extrahovanie vzťahov medzi entitami v texte, ktoré zodpovedajú hranám znalostného grafu
- Výsledok je výsledný znalostný graf – množina faktov v tvare subjekt-predikát-objekt

⟨George H. W. Bush⟩ ⟨stretol sa⟩ ⟨Hamid Karzai⟩

Sémantická reprezentácia – príklad (1)

- Vety vyextrahované z medicínskych článkov

Stress is associated with migraines.
Stress can lead to loss of magnesium.
Calcium channel blockers prevent some migraines
Magnesium is a natural calcium channel blocker.
Spreading cortical depression (SCD) is implicated in some migraines.
High levels of magnesium inhibit SCD.
Migraine patients have high platelet aggregability.
Magnesium can suppress platelet aggregability.

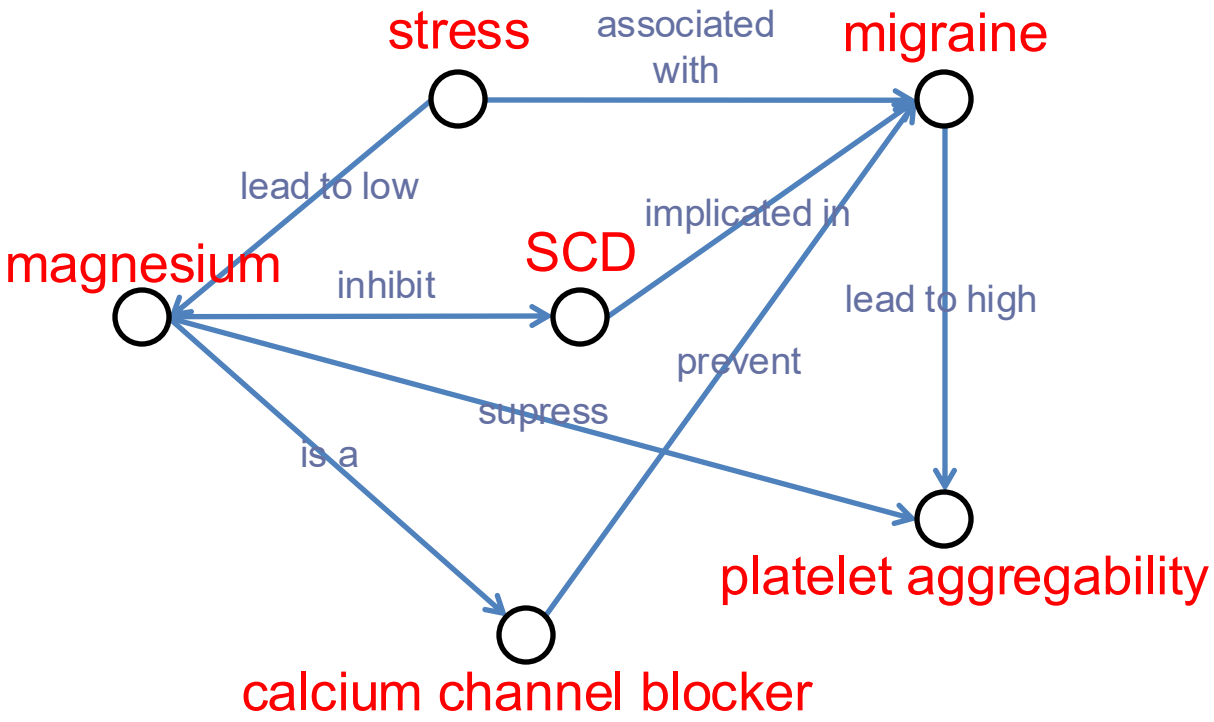
Sémantická reprezentácia – príklad (2)

- Extrahovanie entít

Stress is associated with migraines
Stress can lead to loss of magnesium
Calcium channel blockers prevent some migraines
Magnesium is a natural calcium channel blocker
Spreading cortical depression (SCD) is implicated in some migraines
High levels of magnesium inhibit SCD
Migraine patients have high platelet aggregability
Magnesium can suppress platelet aggregability

Sémantická reprezentácia – príklad (3)

- Extrahovanie relácií



Algoritmy spracovania prirodzeného jazyka

Metódy analýzy prirodzeného jazyka

- Manuálne vytvorenie slovníkov a pravidiel pre extrahovanie informácií nie je dobre škálovateľné a robustné
- Úlohy spracovania prirodzeného jazyka sa dajú transformovať na úlohy strojového učenia pre sekvenčné dáta – text = sekvencia znakov/tokenov
- Vstup je množina textov – textový korpus
- Využíva sa najmä kontrolované, resp. samo-kontrolované učenie

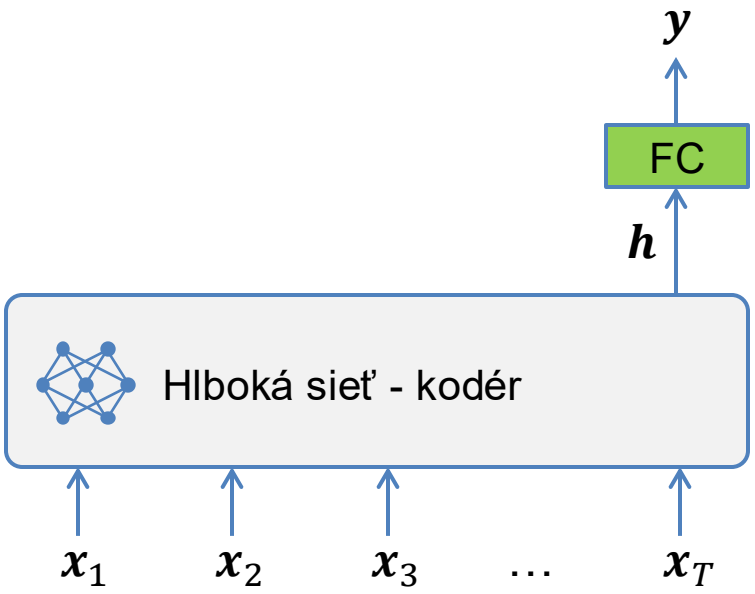
Úlohy strojového učenia pre sekvenčné dáta

- Klasifikácia sekvencií
  - Vstup: sekvencia
  - Výstup: označenie / priradenie triedy
  - Napr. klasifikácia dokumentov, analýza sentimentu
- Mapovanie sekvencií
  - Vstup/výstup: sekvencia
  - Zarovnané sekvencie s rovnakou dĺžkou / sekvencie s rôznou dĺžkou na vstupe a výstupe
  - Napr. jazykové modely, extrahovanie pomenovaných entít, ale aj rozpoznávanie hovorenej reči z audio signálu

Hlboké učenie na sekvenčných dátach

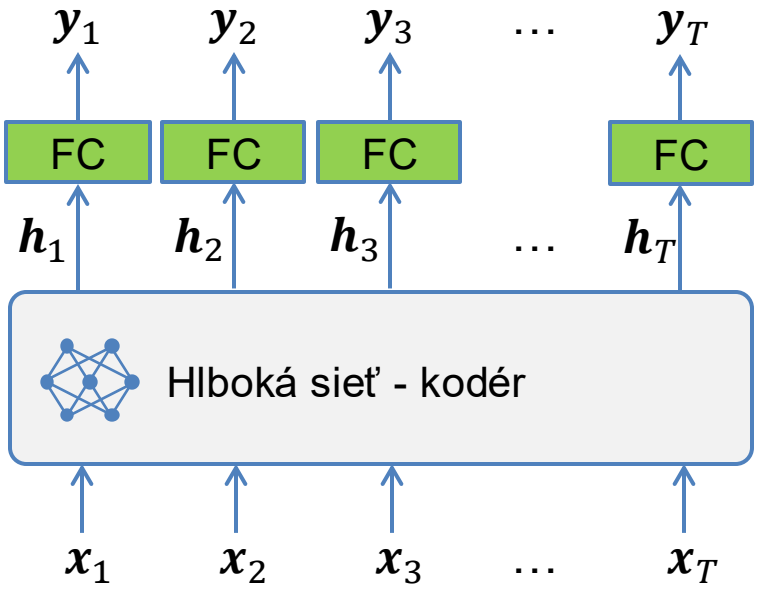
- Využívajú sa najmä metódy hlbokého učenia, ktoré dokážu počas učenia vyextrahovať príznaky potrebné na riešenie danej úlohy na všetkých jazykových úrovniach
- Význam textu a vlastnosti jazykových jednotiek sú zakódované algebricky v podobe číselných vektorov – *embeddings*
- Pre jednoduchšie úlohy sú efektívne aj tradičné metódy strojového učenia + vektorová reprezentácia dokumentov na vstupe

Klasifikácia sekvencií



- Vektor $\boldsymbol{h}$ obsahuje príznaky vyextrahované z celej sekvencie v závislosti na poradí jednotlivých prvkov
- FC - dopredná sieť, ktorá vypočíta výslednú predikciu $\boldsymbol{y}$

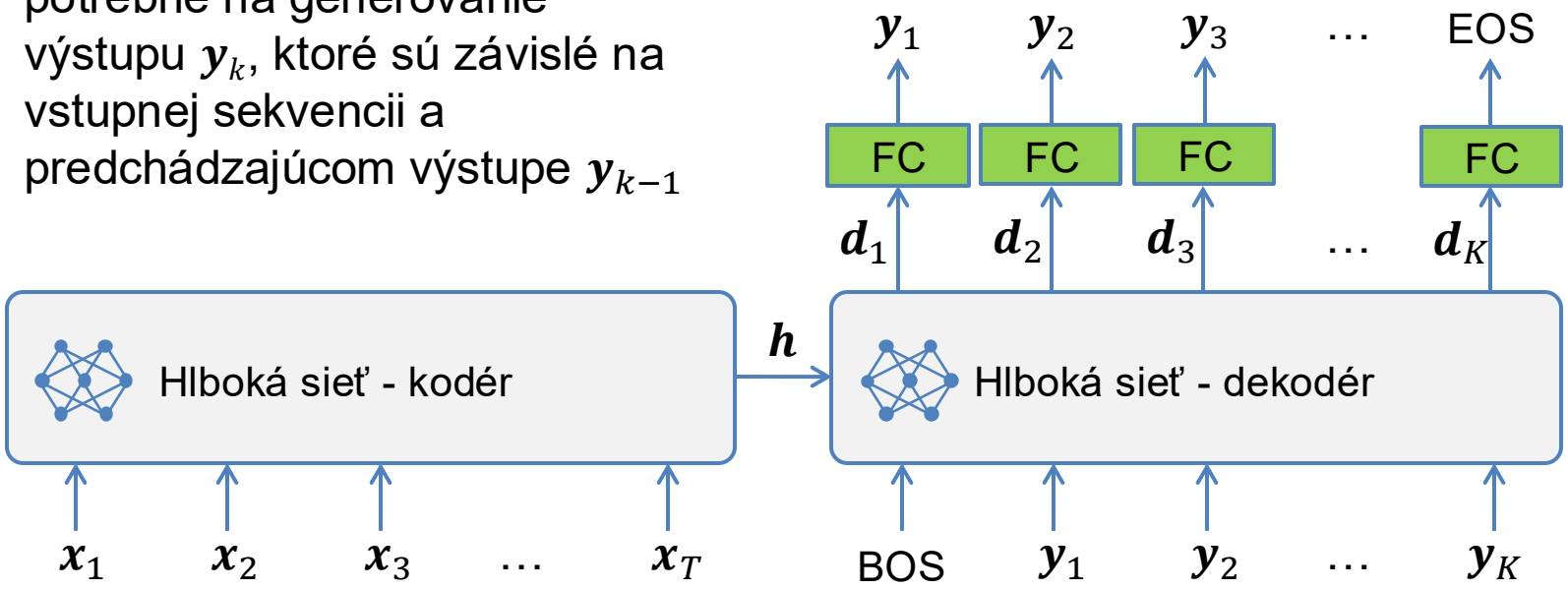
Mapovanie sekvencií rovnakej dĺžky



- Vektory $\boldsymbol{h}_t$ obsahujú vyextrahované príznaky o vstupe $\boldsymbol{x}_t$ v kontexe predchádzajúcich a nasledujúcich vstupov v sekvencii.
- FC je rovnaká pre všetky $\boldsymbol{h}_t$

Mapovanie sekvencií rôznej dĺžky

Vektory $\boldsymbol{d}_k$ obsahujú príznaky potrebné na generovanie výstupu $\boldsymbol{y}_k$, ktoré sú závislé na vstupnej sekvencii a predchádzajúcom výstupe $\boldsymbol{y}_{k-1}$



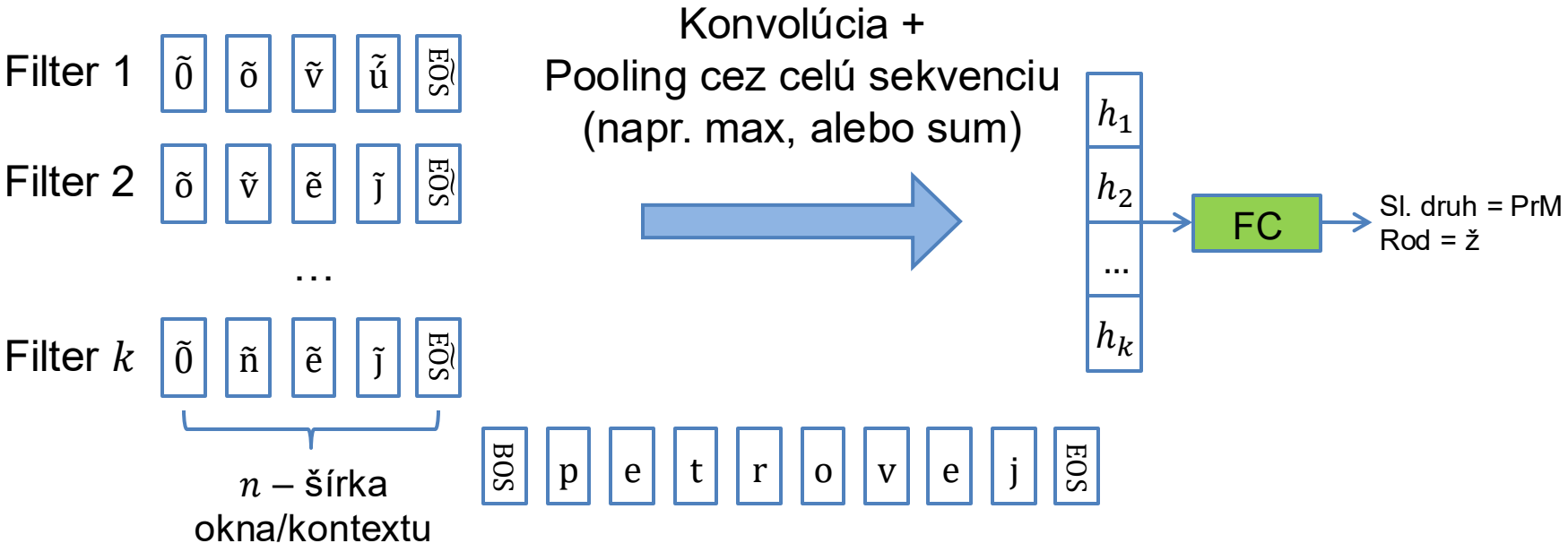
BOS a EOS je špeciálny vstup a výstup označujúci začiatok a koniec generovanej sekvencie

Hlboké siete pre spracovanie prirodzeného jazyka

- Kodér
  - 1D konvolučné siete
  - (Obojsmerné) rekurentné siete
  - Transformery
- Dekodér
  - Rekurentné siete (voliteľne s pozornostným modelom)
  - Transformery

1D konvolučné siete pre texty

- Slúžia na učenie lokálnych príznakov v ohraničenom kontexte (okne)
- Zodpovedajú n-gramom (postupnosti n tokenov/znakov)



Reprezentácia vstupných symbolov

- Pri spracovaní textu je na vstupe postupnosť symbolov – tzn. znakov na najnižšej úrovni, alebo najčastejšie tokenov reprezentujúcich slová
- Každému symbolu je priradený vektor príznakov – *embeddings,* ktorý ho bude reprezentovať na vstupe hlbokej siete
- Vektory môžu byť inicializované náhodne a môžu sa meniť ako parametre počas učenia
  - Naučí sa reprezentácia symbolov potrebná pre riešenie danej špecifickej úlohy
  - Musíme mať dostatočné (označené) trénovacie dáta na výpočet parametrov zavedených pre každý symbol

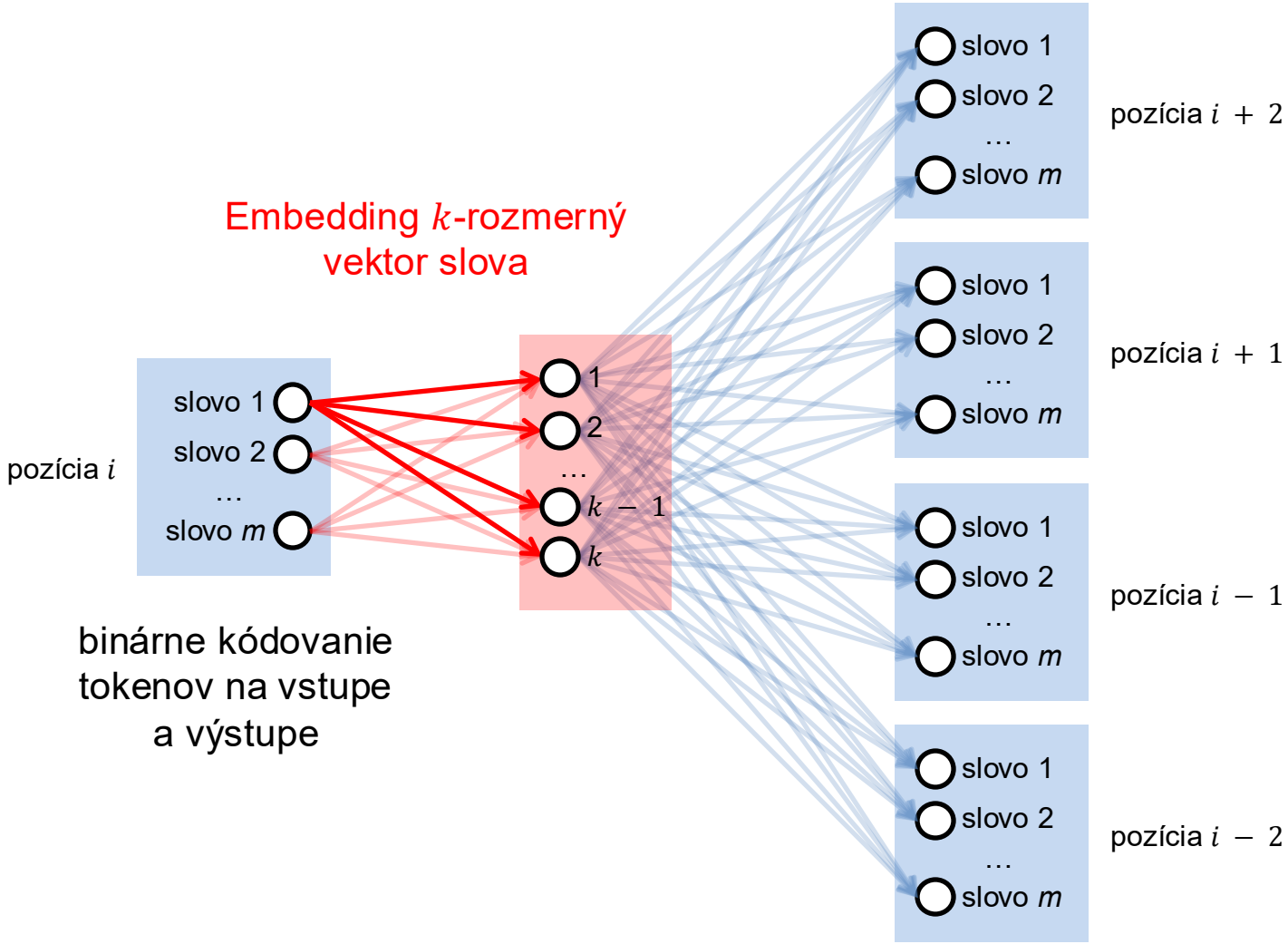
Distribuovaná reprezentácia slov

- Vieme vypočítať vstupné príznaky pre každé slovo bez toho, aby sme použili manuálne anotované dáta?
- Predpoklady: morfologické, syntaktické a sémantické vlastnosti slova sa dajú odvodiť od okolitých slov, ktoré sa spolu vyskytujú s daným slovom
- Slová sú si podobné, ak sa vyskytujú v podobnom kontexte
- Kontext vieme priamo získať aj bez manuálneho označenia dát z rozsiahleho korpusu textov

Word2Vec (skip-gram) model (1)

- Naučíme model, ktorý bude pre každé slovo predpovedať, aké slová môžu nasledovať alebo predchádzať danému slovu
- V parametroch modelu budú zakódované *embedding* vektory slova/tokenu, ktoré potom môžeme použiť ako vstup pre špecifické úlohy spracovania prirodzeného jazyka

Word2Vec (skip-gram) model (2)



GloVe – globálne vektory (1)

- Metóda vychádza z globálnej matice spolu-výskytov slov v celom korpuse (tzn. pre každé slovo i a j vypočítame $X_{i,j}$, koľko krát sa slovo j vyskytlo v okolí slova i)
- Optimalizačná úloha: vytvoríme lineárny model medzi vektormi $\boldsymbol{w}_i$ a $\tilde{\boldsymbol{w}}_j$ tak aby sa jeho predikovaná hodnota čo najviac podobala logaritmu pravdepodobnosti spolu-výskytu slova i a j, tzn.

$$\boldsymbol{w}_i\tilde{\boldsymbol{w}}_j + b_i + \tilde{b}_j \approx lnP_{i,j} = ln\left(\frac{X_{i,j}}{X_i}\right)$$

kde X_i je počet slov, ktoré sa celkovo vyskytli v okolí slova i (tzn. 2 $\times$ šírka kontextu $\times$ počet výskytov slova i)

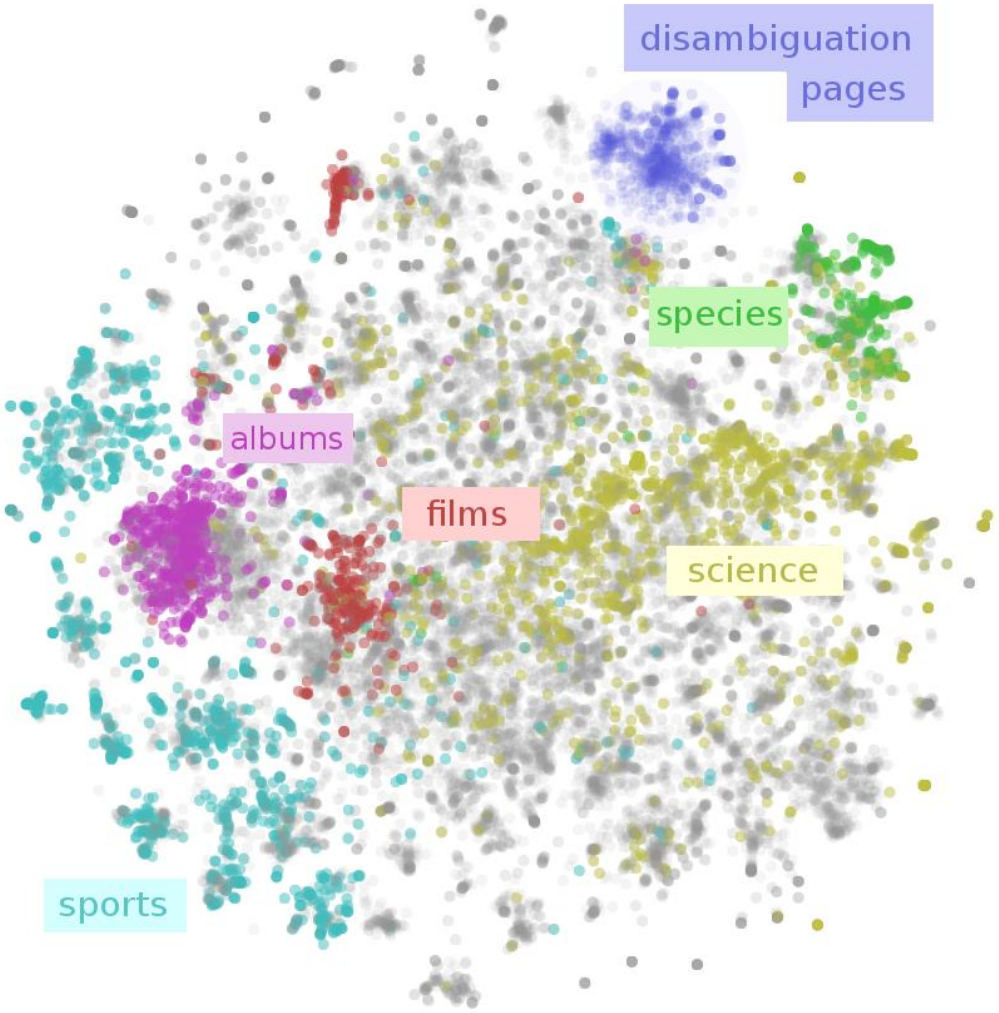
GloVe – globálne vektory (2)

- Ku každému slovu získame dvojicu vektorov $\boldsymbol{w}_i$ a $\tilde{\boldsymbol{w}}_i$, výsledný *embedding* slova je ich kombináciou $\boldsymbol{w}_i + \tilde{\boldsymbol{w}}_i$
- Interpretácia: slová majú podobné vektory (skalárny súčin je veľké číslo) ak sa často vyskytujú v podobnom kontexte, tzn. majú podobnú distribúciu $P_{i,k} \approx P_{j,k}$

Vlastnosti distribuovanej reprezentácie slov

- Word2Vec a GloVe modely dokážu zachytiť niektoré sémantické vzťahy medzi slovami
- Synonymá sú často namapované na podobné vektory
- Empiricky sa ukazuje, že je možné použiť aritmetické operátory na rozšírenie a kombinovanie významu slov, napr. *muž* – *žena* ≈ *brat* – *sestra*, *jazyk* + *USA* ≈ *angličtina*, a pod.

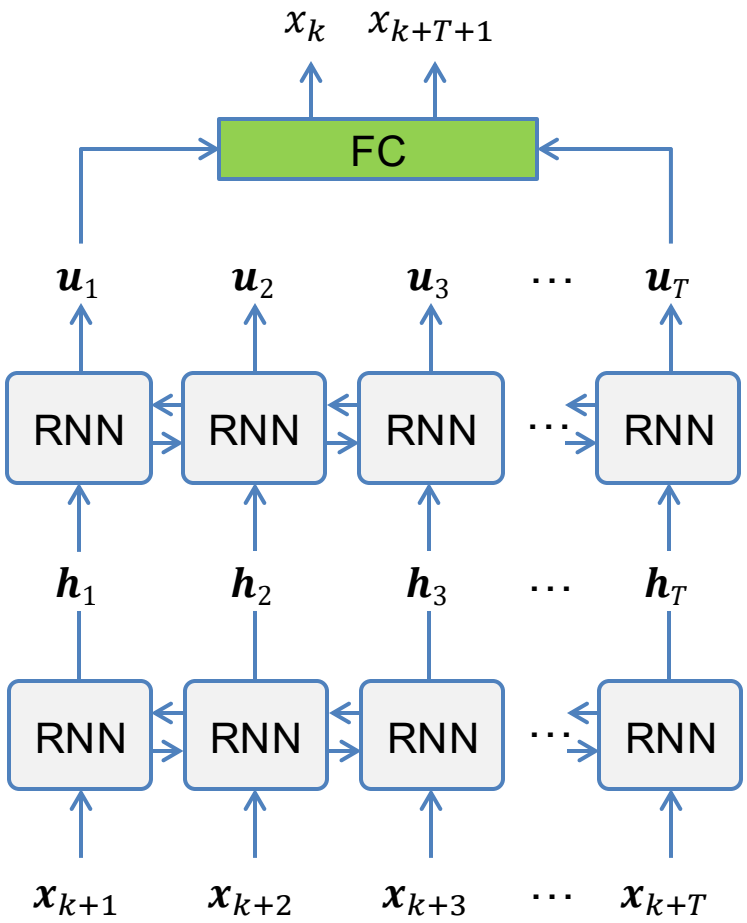
Vizualizácia *embedding* vektorov podľa Wikipedie



ELMo (1)

- Word2Vec a GloVE nedokážu rozlíšiť slová, ktoré majú viacero významov podľa aktuálneho kontextu – jednému slovu je vždy priradený ten istý statický vektor
- Pri ELMo sa *embedding* vektor pre slovo vypočíta pomocou jazykového modelu naučenom na rozsiahlom korpuse – pre rôzne výskyty toho istého slova môže byť odlišný podľa kontextu v texte
- Ako model je možné použiť napr. viacvrstvovú rekurentnú obojsmernú sieť, alebo transformer

ELMo (2)



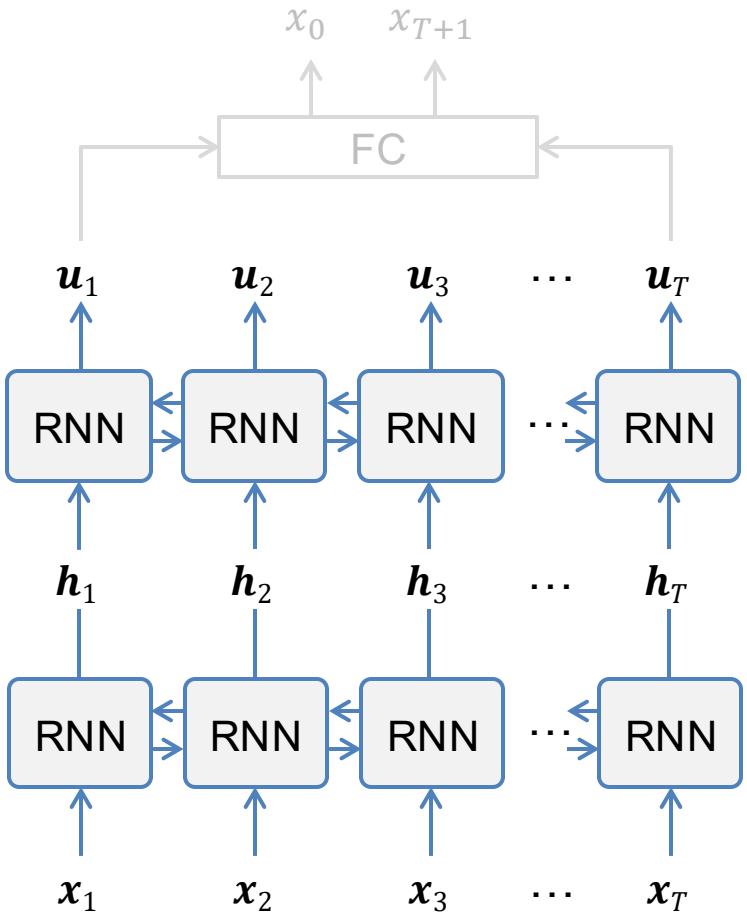
Najprv naučíme jazykový model na rozsiahlom korpuse

Pre obojsmernú rekurentnú sieť:

- dopredná RNN sa učí predikovať nasledujúce slovo v korpuse
- spätná RNN predchádzajúce slovo

Po naučení jazykového modelu ho použijeme na predikciu pre vstupný text, pre ktorý chceme vypočítať *embedding* reprezentáciu

ELMo (3)



Pre každé slovo dostaneme pôvodný "statický" embedding vektor $\boldsymbol{x}_i$ a stavy rekurentných sietí $\boldsymbol{h}_i$ a $\boldsymbol{u}_i$

Výsledný *embedding* vektor vypočítame lineárnou kombináciou

$$\boldsymbol{v}_i = [\boldsymbol{x}_i, \boldsymbol{h}_i, \boldsymbol{u}_i]\mathbf{X}$$

kde $\mathbf{X}$ je projekčná matica, ktorú je možné doučiť pre konkrétnu úlohu