Lexikální disambiguace a distribuční sémantické modely

Martin Holub

Několik poznámek o souvislostech

- ***lexikální disambiguace***
- ***selekčních preferencí***
- ***distribučních modelů***
- ***a řídkých dat***

Co je Distribuční Sémantický Model?

- DSM dynamicky buduje sémantické reprezentace slov – na základě analýzy kontextů v korpusu

- Technicky jde zpravidla o konstrukci vektorů popisujících (charakteristiký) kontext(y) slov (vzniká vektorový prostor, a lze využít dobře definované algebraické operace)

- ÚČEL: DSM je obvykle primárně využíván k odhadu/výpočtu sémantické podobnosti slov

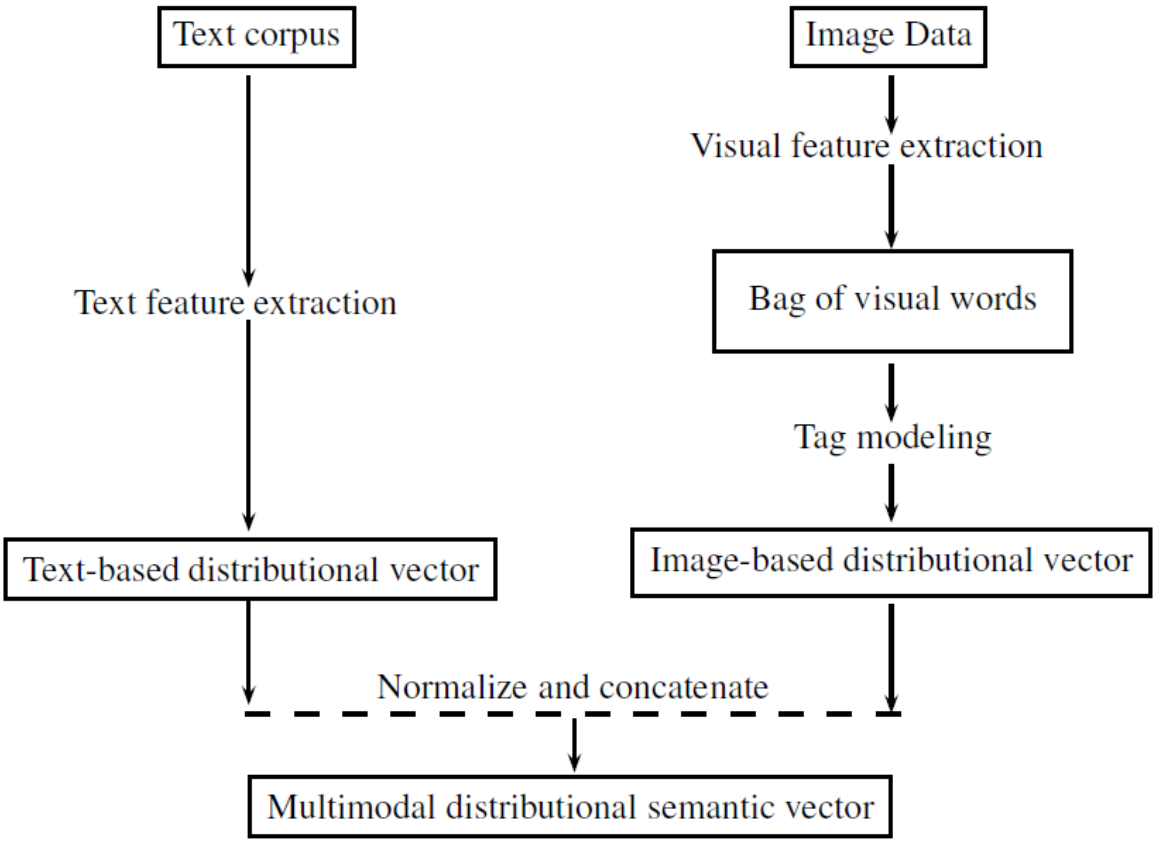
- Výhody DSM oproti jiným přístupům
  - *flexibilita* (lze budovat odlišné modely v závislosti na doméně)
  - použitelné pro *unsupervised learning* -- nevyžaduje ruční anotace a umožňuje zpracovat veliké korpusy

- CLUSTERING v rámci DSM může odhalit
  - jak *sémantické třídy* složené z různých slov
  - tak i naopak shluky sémanticky podobných kontextů jednoho slova → hypoteticky *sémantické kategorie*

Kombinace multimodálních dat v DSM

- náš student Giang Tran (+ E. Bruni, M. Baroni, R. Bernardi) kombinuje distribuční model vytvořený z textů a distribuční model vytvořený z obrázků, využívá data z ESP databáze

- relativně úspěšné při evaluaci pomocí několika různých standardních testovacích ručně anotovaných datových sad (sémantická podobnost)



Selekční preference slovesa

Popisují znalost o tom, co je/může být v kontextu slovesa jako jeho argument

- formálně je "síla" selekční preference buď pravděpodobnost, nebo míra náležení do fuzzy množiny

- "síla" preference vyjadřuje, zda je slovo v daném kontextu
  - typické/charakteristické
  - použitelné
  - vzácné až prakticky nemožné

- selekční preference považujeme za jeden z klíčů k úspěšnému rozpoznávání PDEV-patternů, protože
  - patterny jsou de facto pomocí selekčních preferencí definovány
  - také naše (lidská) zkušenost ukazuje, že se při rozpoznávání významu do značné míry řídíme typem argumentů (což byla Hansova motivace pro definování sémantických kategorií pomocí patternů)

- ústřední překážkou pro identifikaci lexikálního obsazení sémantických typů použitých v definicích patternů je nedostatek dat (*data sparsity*)
  - publikované práce ukazují, že k částečnému překonání tohoto problému může pomoci DSM – ten totiž ukáže sémanticky podobná slova neviděná v trénovacích datech