Exploring Three Way Contexts for Word Sense Discrimination

Tim Van de Cruys

University of Groningen

CoSMo 2007
Beyond Words and Documents
August 21, 2007

RuG

Semantic similarity

- Most work on semantic similarity relies on the DISTRIBUTIONAL HYPOTHESIS (Harris 1954)
- Take a word and its contexts:
  - fresh *smørrebrød*
  - tasty *smørrebrød*
  - a portion of *smørrebrød*
  - *smørrebrød* topped with herring
- By looking at a word's context, one can infer its meaning

RuG

Semantic similarity

- Most work on semantic similarity relies on the DISTRIBUTIONAL HYPOTHESIS (Harris 1954)
- Take a word and its contexts:
  - fresh *smørrebrød*
  - tasty *smørrebrød*
  - a portion of *smørrebrød* $\Rightarrow$ **FOOD**
  - *smørrebrød* topped with herring
- By looking at a word's context, one can infer its meaning

RuG

Two kinds of context

1. 'Bag of words' context
   - a window around the word is used as context
   - e.g. a fixed numbers of words, the paragraph in which a word appears, ...
   - often used with some form of dimensionality reduction
2. Syntactic context
   - a corpus is parsed, dependency triples are extracted
   - e.g. <apple, obj, eat>, <apple, adj, red>
   - typically does not use any form of dimensionality reduction

RuG

Ambiguity

- **Problem**: ambiguity
  - Compare:
    a trendy bar
    $\leftrightarrow$ *an iron bar*
    $\leftrightarrow$ *today's air pressure: 1.013 bar*
  - Different meanings, but they are considered the same entity by a naive algorithm
- Main research question: can 'bag of words' context and syntactic context be combined to differentiate between various senses of a word?

RuG

Distributional Similarity

1. Latent Semantic Analysis (LSA)
   - the application of a mathematical/statistical technique (SVD) in order to acquire semantics from plain text
   - LSA tries to find 'latent semantic dimensions' according to which words can be identified
   - Criticized for using an objective function not appropriate for textual data
   - Subsequent dimensionality reduction algorithms (Probabilistic Latent Semantic Analysis, Non-Negative Matrix Factorization) remedy the flaws of LSA
2. Syntactic context [Lin 1998]

RuG

Word Sense Discrimination

- Context-group discrimination [Schütze 1998]
  - Clustering the context of ambiguous words
  - Second-order co-occurrence: the contexts of words are similar if the words they co-occur with are similar
  - bag of words context
- Clustering by Committee [Pantel and Lin 2002]
  - First, find tight, unambiguous clusters
  - Next, assign words to clusters, and strip off features associated with a particular cluster
  - syntactic context

RuG

Technique 1/2

- Given a non-negative matrix V, find non-negative matrix factors W and H such that:

$$V_{nxm} \approx W_{nxr} H_{rxm} \quad (1)$$

- Choosing $r \ll n, m$ reduces data
- Constraint on factorization: all values in three matrices need to be *non-negative values* (≥ 0)
- Constraint brings about a *parts-based* representation: only additive, no subtractive relations are allowed

RuG

Technique 2/2

- Different kinds of NMF's that minimize different cost functions:
  - Square of Euclidean distance (L1-norm)
  - Kullback-Leibler Divergence (L2-norm)
    $\Rightarrow$ better suited for language phenomena
- To find NMF is to minimize $D(V \| WH)$ with respect to W and H, subject to the constraints $W, H \geq 0$
- This can be done with *update rules*

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} \frac{V_{i\mu}}{(WH)_{i\mu}}}{\sum_v H_{av}} \qquad (2)$$

- these update rules converge to a *local optimum* in the minimization of KL divergence

RuG

Results

- Context vectors (10k nouns × 2k co-occurring nouns) extracted from CLEF corpus
- NMF is able to capture semantic dimensions (much more obvious and clear than LSA)
- Examples:
  - *bus* 'bus', *taxi* 'taxi', *trein* 'train', *halte* 'stop', *reiziger* 'traveler', *perron* 'platform', *tram* 'tram', *station* 'station', *chauffeur* 'driver', *passagier* 'passenger'
  - *bouillon* 'broth', *slagroom* 'cream', *ui* 'onion', *eierdooier* 'egg yolk', *laurierblad* 'bay leaf', *zout* 'salt', *deciliter* 'decilitre', *boter* 'butter', *bleekselderij* 'celery', *saus* 'sauce'

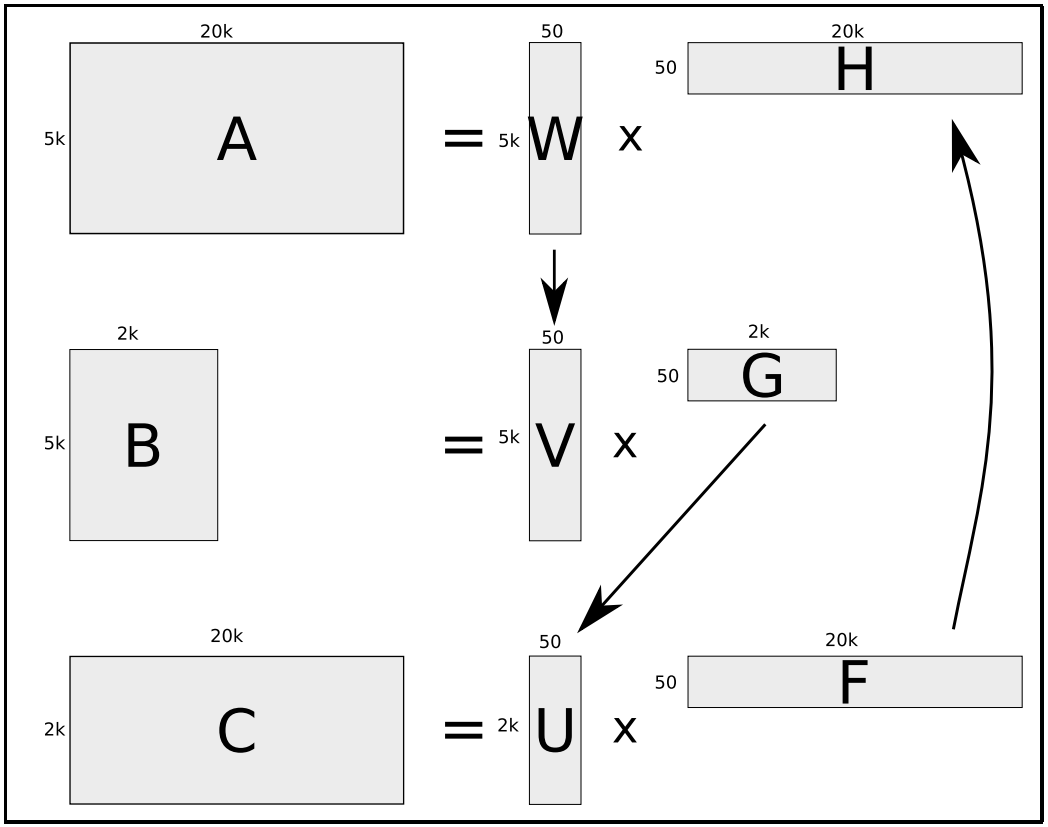
RuG

Methodology

- Goal: classification of nouns according to both 'bag of words' context and syntactic context
- $\Rightarrow$ Construct three matrices capturing co-occurrence frequencies for each mode
  - nouns cross-classified by dependency relations
  - nouns cross-classified by (bag of words) context words
  - dependency relations cross-classified by context words
- $\Rightarrow$ Apply NMF to matrices, but interleave the process
- Result of former factorization is used to initialize factorization of the next one

RuG

Graphical Representation



Sense subtraction

- 'switch off' one dimension of an ambiguous word to reveal other possible senses
- From matrix W, we know which dimensions are the most important for a certain word
- Matrix H gives the importance of each dependency relation given a dimension
- 'subtract' dependency relations that are responsible for a given dimension from the original noun vector
  - $\overrightarrow{v}_{new} = \overrightarrow{v}_{orig}(\overrightarrow{1} - \overrightarrow{h}_{dim})$
  - each dependency relation is multiplied by a scaling factor, according to the load of the feature on the subtracted dimensions

RuG

Experimental Design

- Approach applied to Dutch, using CLEF corpus (Dutch newspaper texts '94-'95)
- Corpus parsed with Dutch dependency parser ALPINO
- three matrices constructed with:
  - 5k nouns $\times$ 20k dependency relations
  - 5k nouns $\times$ 2k context words
  - 20k dependency relations $\times$ 2k context words
- Factorization to 50 dimensions

RuG

Example dimension: transport

1. **nouns**: *auto* 'car', *wagen* 'car', *tram* 'tram', *motor* 'motorbike', *bus* 'bus', *metro* 'subway', *automobilist* 'driver', *trein* 'trein', *stuur* 'steering wheel', *chauffeur* 'driver'
2. **context words**: *auto* 'car', *trein* 'train', *motor* 'motorbike', *bus* 'bus', *rij* 'drive', *chauffeur* 'driver', *fiets* 'bike', *reiziger* 'reiziger', *passagier* 'passenger', *vervoer* 'transport'
3. **dependency relations**: viertraps$_{adj}$ 'four pedal', verplaats_met$_{obj}$ 'move with', toeter$_{adj}$ 'honk', tank_in_houd$_{obj}$ [parsing error], tank$_{subj}$ 'refuel', tank$_{obj}$ 'refuel', rij_voorbij$_{subj}$ 'pass by', rij_voorbij$_{adj}$ 'pass by', rij_af$_{subj}$ 'drive off', peperduur$_{adj}$ 'very expensive'

RuG

Pop

pop music $\leftrightarrow$ *doll*

1. *pop*, *rock*, *jazz*, *meubilair* 'furniture', *popmuziek* 'pop music', *heks* 'witch', *speelgoed* 'toy', *kast* 'cupboard', *servies* '[tea] service', *vraagteken* 'question mark'
2. *pop*, *meubilair* 'furniture', *speelgoed* 'toy', *kast* 'cupboard', *servies* '[tea] service', *heks* 'witch', *vraagteken* 'question mark' *sieraad* 'jewel', *sculptuur* 'sculpture', *schoen* 'shoe'
3. *pop*, *rock*, *jazz*, *popmuziek* 'pop music', *heks* 'witch', *danseres* 'dancer', *servies* '[tea] service', *kopje* 'cup', *house* 'house music', *aap* 'monkey'

RuG

Barcelona

Spanish city $\leftrightarrow$ *Spanish football club*

1. *Barcelona*, *Arsenal*, *Inter*, *Juventus*, *Vitesse*, *Milaan* 'Milan', *Madrid*, *Parijs* 'Paris', *Wenen* 'Vienna', *München* 'Munich'
2. *Barcelona*, *Milaan* 'Milan', *München* 'Munich', *Wenen* 'Vienna', *Madrid*, *Parijs* 'Paris', *Bonn*, *Praag* 'Prague', *Berlijn* 'Berlin', *Londen* 'London'
3. *Barcelona*, *Arsenal*, *Inter*, *Juventus*, *Vitesse*, *Parma*, *Anderlecht*, *PSV*, *Feyenoord*, *Ajax*

RuG

Conclusion

- Combining bag of words data and syntactic data is useful
  - bag of words data (combined with NMF) puts its finger on topical dimensions
  - syntactic data is particularly good at finding similar words
  - use of three way data allows one to determine which topical dimension(s) are responsible for a certain sense
  - and adapt the (syntactic) feature vector accordingly
  - by subtracting one sense to discover less dominant senses
- Exploratory: more research, an elaborated framework and in-depth evaluation needed

RuG

Future Work

- A sound statistical framework
  - Instability of the NMF
    $\leftrightarrow$ Latent Dirichlet Allocation, Multinomial PCA
  - 'ad hoc' subtraction of topical dimensions
    $\leftrightarrow$ statistical modeling of the topical dimensions
- Proper evaluation, comparison with other WSD algorithms
- Embed in a clustering approach
  - determine which dimensions are important for a given cluster
  - subtract these dimensions from the individual words to see if other sense emerge

RuG