NLP Technologies for Cognitive Computing
Lecture 3: Word Senses

Devdatt Dubhashi

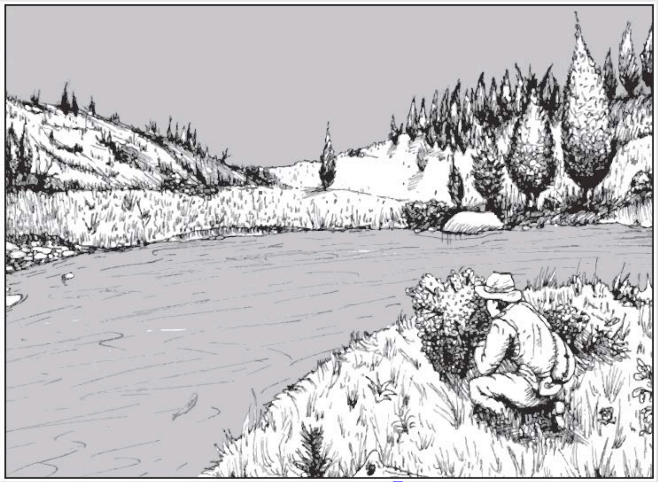
LAB

(Machine Learning. Algorithms, Computational Biology)

Computer Science and Engineering

Chalmers

Why Language is difficult ..



polysemous

synonymous

Concept Layer

Lexical Layer

He sat on the river bank and counted his dough.

She went to the bank and took out some money.


Geology
Name the three types of rock.
1. Classic
2. Punk
3. Hard

WSI and WSD

- Word sense induction (WSI): the automatic identification of the senses of a word (i.e. meanings).
- Word sense disambiguation (WSD): identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings.

WSI and WSD: Two approaches

Knowledge Based

- Detailed
- Interpretable'
- High quality

Corpus Based/Data Driven

- Greater coverage
- Novel senses (previously unknown)

WORD SENSE INDUCTION

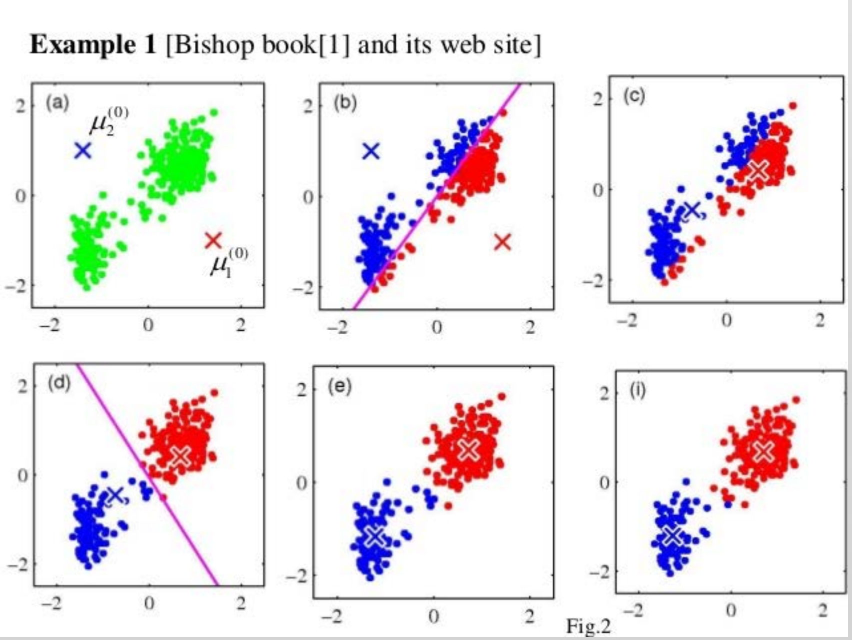
Context and Target Vectors in word2vec

- Assign to each word *w*, a target vector $\boldsymbol{u}_w$ and a context vector $\boldsymbol{v}_w$ in $\boldsymbol{R}^d$
- Target vector represents the meaning of the word
- Context vector represents the word when it is used in the context of another word.

Key Idea: Cluster contexts

- For each word, look at all the contexts in which it occurs in the corpus
- Clustering those context vectors should give information about its different senses – the Distributional Hypothesis!

K-Means Clustering



- Initialze k centers $c_1 \ldots c_k$
- Repeat until convergence:
  - Assign each point x_j to its closest center, call the set assigned to c_i, C_i
  - Recompute the centers: $c_i = \frac{1}{|C_i|}\sum_{j \in C_i} x_j$

Centers: How many and which?

- Performance of k Means depends very strongly on :
  - Where the initial centers are chosen at the start
  - How many centers are chosen, the parameter k.

How to Initialize I

- Pick k random points
- Pick k points at random from input points
- Assign points at random to k groups, and take their centroids as initial centers
- Pick first center at random, take next center as far away from first, take next as far away from first two …

K Means ++

- Let $d(x, C) := min_{c \in C}\, d(x, c)$
- Start with C_1 containing a single point picked from input uniformly at random
- For $k \geq 2$, let

$$c^* = \text{argmax}\{\frac{d^2(x, C_{k-1})}{\sum_y d^2(y, C_{k-1})}\}$$

  - $C_k = C_{k-1} \cup \{c^*\}$
  - D. Arthur, S. Vassilvitskii (2007): $O(\log n)$ approximation to the optimal.
  - Disadvantage: needs k passes through input, running time $O(nkd)$

$AFK - MC^2$ (NIPS 2016)

MCMC approach to sampling from the target distribution.

Algorithm 1 ASSUMPTION-FREE K-MC^2(AFK-MC^2)

Require: Data set $\mathcal{X}$, # of centers k, chain length m

```
    // Preprocessing step
    c_1 ← Point uniformly sampled from 𝒳
    for all x ∈ 𝒳 do
        q(x) ← ½ d(x, c_1)^2 / Σ_{x'∈𝒳} d(x', c_1)^2 + 1/(2n)
    // Main loop
    C_1 ← {c_1}
    for i = 2, 3, ..., k do
        x ← Point sampled from 𝒳 using q(x)
        d_x ← d(x, C_{i-1})^2
        for j = 2, 3, ..., m do
            y ← Point sampled from 𝒳 using q(y)
            d_y ← d(y, C_{i-1})^2
            if (d_y q(x)) / (d_x q(y)) > Unif(0, 1) then x ← y, d_x ← d_y
        C_i ← C_{i-1} ∪ {x}
    return C_k
```

$$q(x \mid c_1) = \frac{1}{2} \underbrace{\frac{d(x, c_1)^2}{\sum_{x' \in \mathcal{X}} d(x', c_1)^2}}_{\text{(A)}} + \frac{1}{2} \underbrace{\frac{1}{|\mathcal{X}|}}_{\text{(B)}} .$$

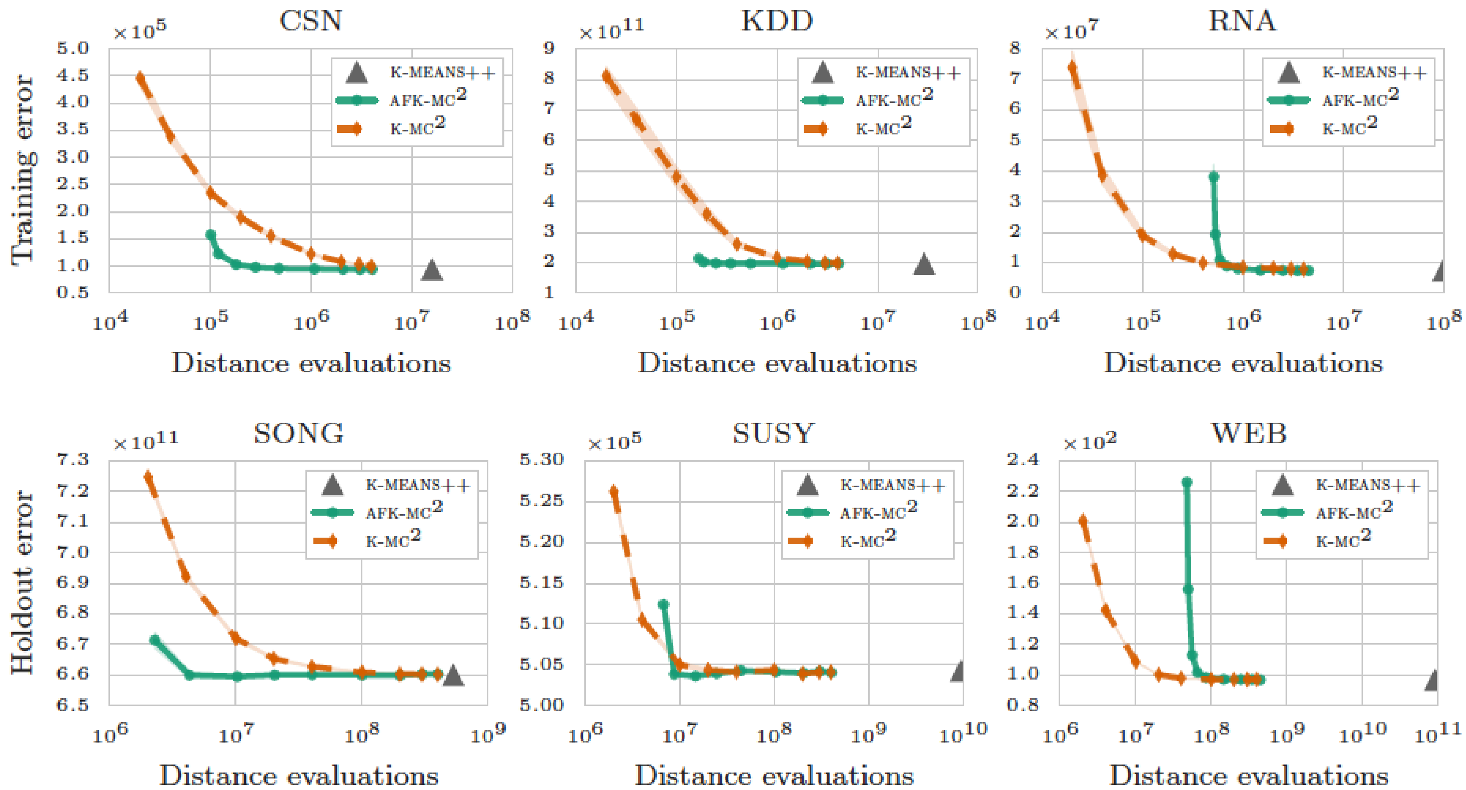


Figure 2: Quantization error in relation to the number of distance evaluations for ASSUMPTION-FREE K-MC^2, K-MC^2 and `k-means++`. ASSUMPTION-FREE K-MC^2 provides a speedup of up to several orders of magnitude compared to `k-means++`. Results are averaged across 200 iterations and shaded areas denote 95% confidence intervals.

Non-parametric clustering: k?

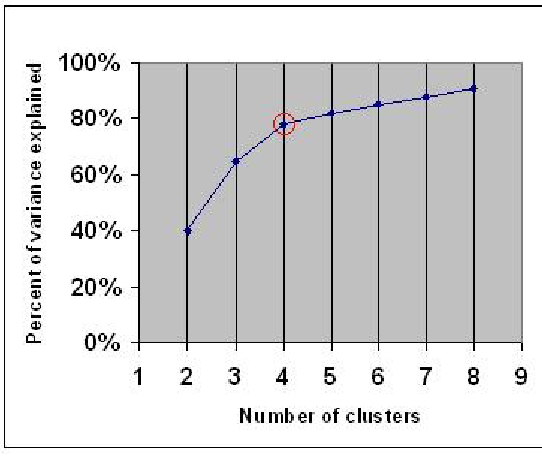
- Intra-cluster *variance*:

$$W_k := \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \mathbf{c}_i)^2$$

- $W = \sum_k W_k$
- Heuristic: Choose k to minimize W
- Elbow Heuristic
- Gap-Statistic: Choose minimum k

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}.$$

$$\mathrm{Gap}_n(k) = E_n^*\{\log W_k\} - \log W_k.$$



Non-parametric clustering via convex relaxations: SON

$$\min_{\mu} \sum_{i} \|x_i - \mu_i\|^2 + \lambda \sum_{i<j} \|\mu_i - \mu_j\|_2$$

Goodness of fit regularization

Quiz

- If we ignore the second (regularization) term, what is the optimal solution to the problem?

SON: Sparsity inducing norm

- The regularization term is a group norm penalty: it will force $\mu_i = \mu_j$ many centroid pairs (μ_i, μ_j).
- Thus for appropriate λ the right number of clusters will be identified automatically tailored to the data without user intervention.

SON versus k-means

- No need to specify k
- Can be used incrementally: as data comes in, the number of clusters adjusts automatically.
- Convex problem, hence unique optimum and no problems of initialization etc.
- Solved efficiently for large data sets by stochastic proximal gradient descent.

SON properties

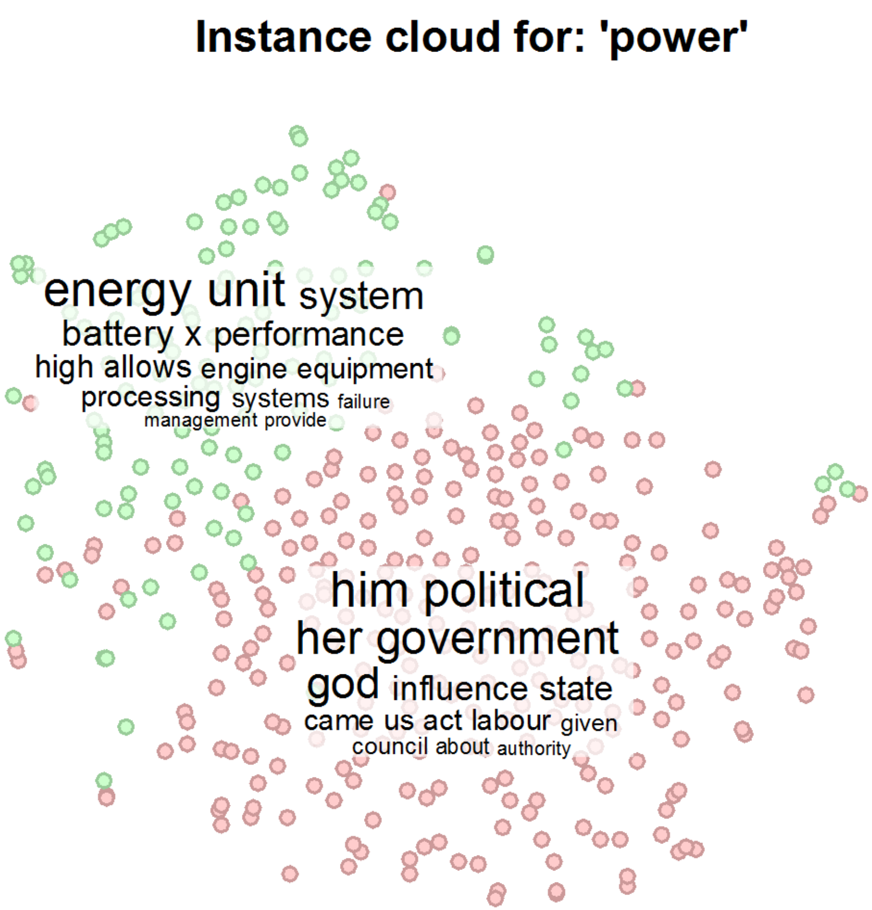
- If the input data has a clear cluster structure:
  - a mixture of well separated Gaussians
  - a stochastic block model

Then the SON optimization problem is guaranteed to recover the clusters perfectly.

ICE Clustering of Context Vectors

- For each word occurrence w, form the weighted centroid of its context vectors: $c_w \ = \ \sum_{w' \in N(w)} \alpha_{w,w'} \, v_{w'}$
- $\alpha_{w.w'} \ = \ \sigma(u_w v_{w'})$
- Also use a triangular context window to give higher weight to words closer to target.
- Now apply k-means to the centroid vectors c_w

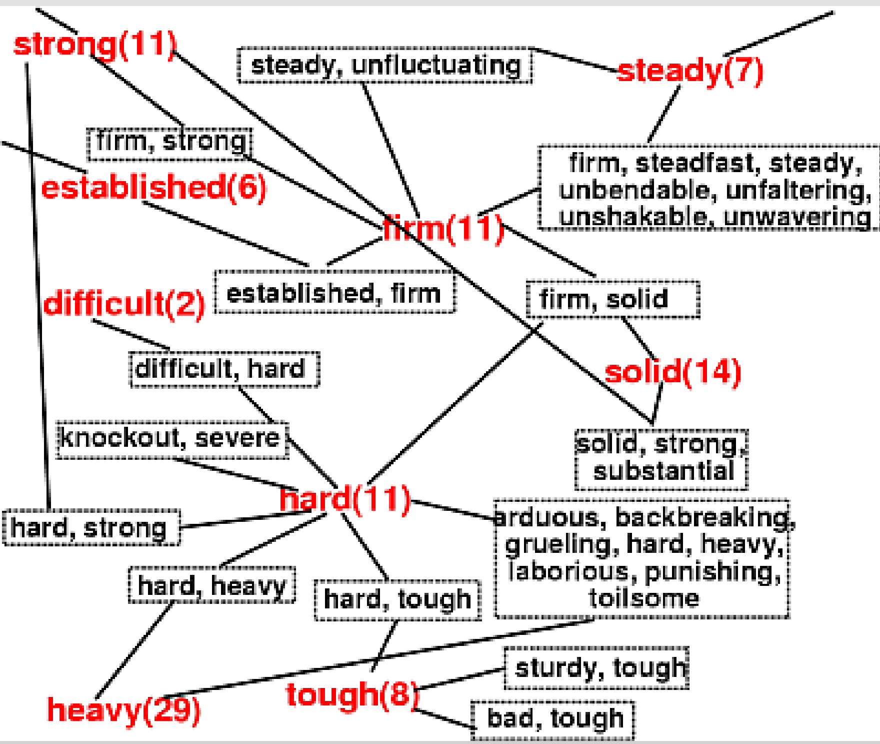
Word sense induction



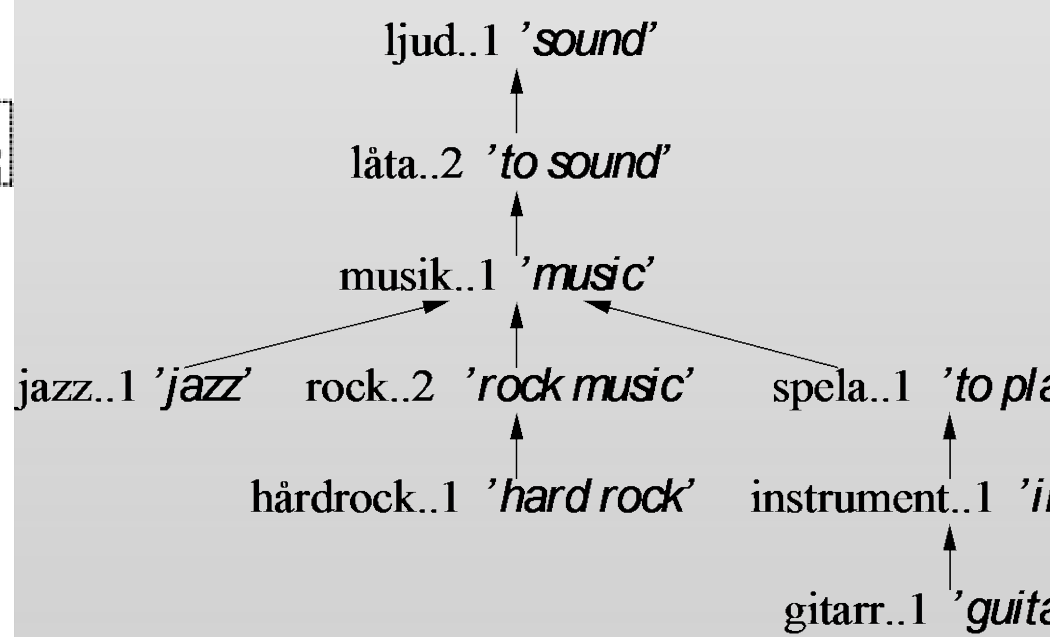
M. Kageback, F. Johansson et al, "Neural context embeddings for automatic discovery of word senses", (NAACL 2015 workshop on Vector Space Modeling for NLP)

Semantic Ontologies

WordNet



SALDO



Fitting word vectors to ontologies

- Split each word vector into vectors corresponding to its different *senses*
- Assign the vector for a particular sense to the corresponding node of the semantic network.

Two Objectives

- For each word vector $u_{w,}$ compute vectors $u_{w1,}$….,u_{wr} corresonding to its r different senses in the network
- Minimize reconstruction error:

$$\left\|\mathbf{u}_w - \alpha_1 \mathbf{u}_{w_1} - \cdots - \alpha_s \mathbf{u}_{w_s}\right\|^2$$

- Maximize fit to network: CBOW model

$$\sum_{w' \in N_G(w)} \log \frac{1}{1 + e^{-\mathbf{u}_{w_i} \cdot \mathbf{v}_{w'}}}$$

Overall Optimization Problem

$$\min \quad \left\|\mathbf{u}_w - \alpha_1 \mathbf{u}_{w_1} - \cdots - \alpha_s \mathbf{u}_{w_s}\right\|^2$$

$$- \sum_{w' \in N_G(w)} \log \frac{1}{1 + e^{-\mathbf{u}_{w_i} \cdot \mathbf{v}_{w'}}}$$

http://demo.spraakdata.gu.se/richard/scouse/

- What kind of optimization problem is this?
- What method would you use to solve it?

WORD SENSE DISAMBIGUATION

WSD: Use context

- Given an occurrence of a word in a text, to disambiguate which sense is being used …
- … use the surrounding context.
- Use sense vectors and context vectors from word2vec!

CBOW approach

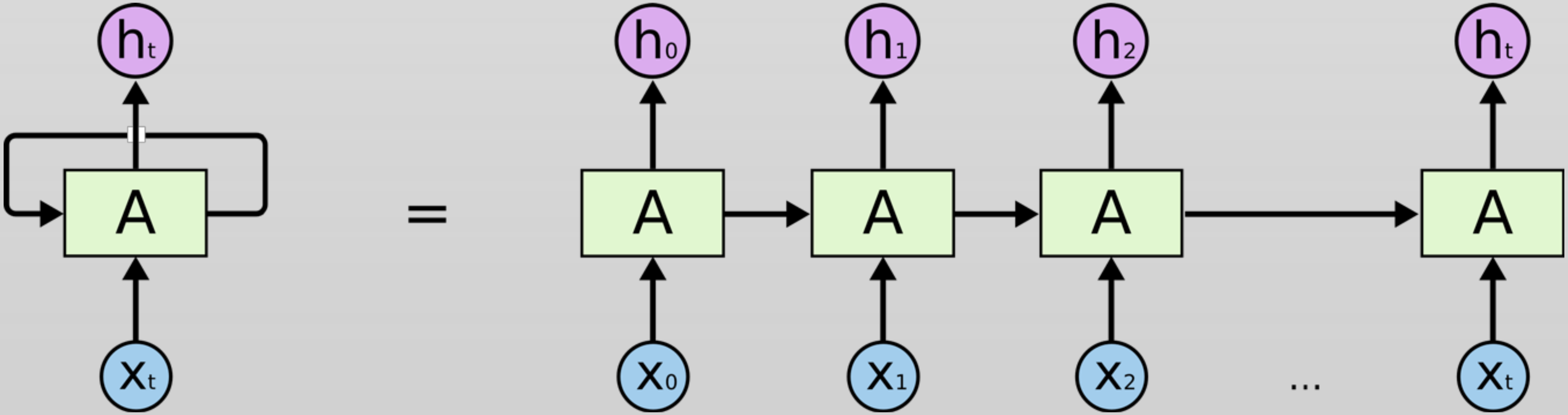
$$P(w_s \mid w') = \frac{1}{1 + e^{-\mathbf{u}_{w_s} \cdot \mathbf{v}_{w'}}}$$

$$P(w_s \mid w'_1, w'_2, \cdots, w'_n) = P(w_s \mid w'_1) P(w_s \mid w'_2) \cdots P(w_s \mid w'_n)$$

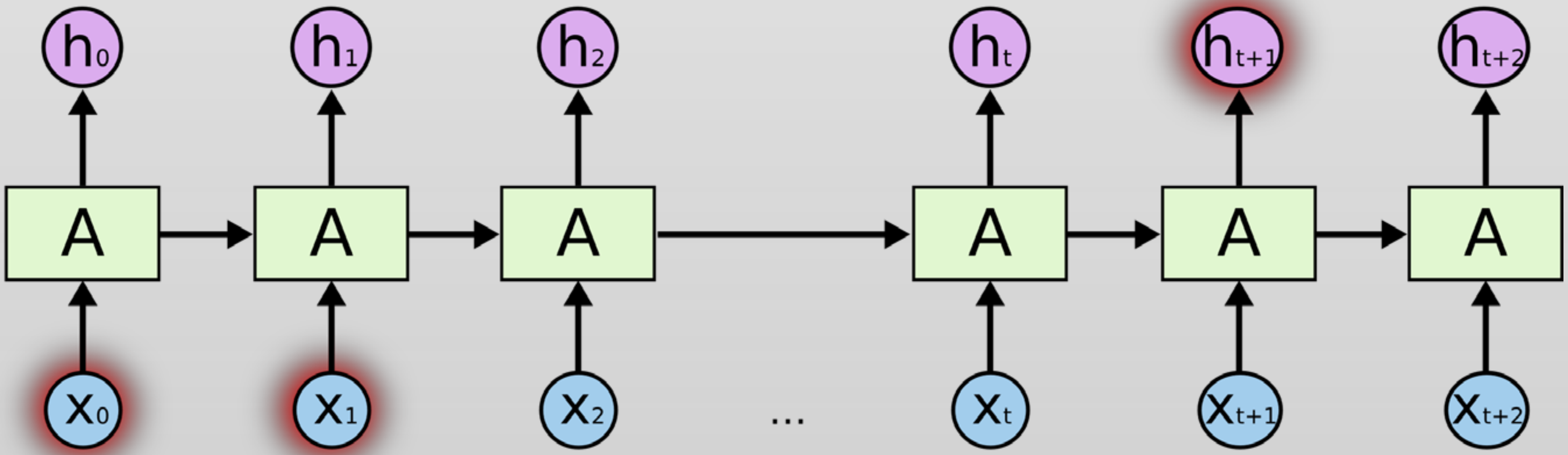
$$argmax_s \log P(w_s \mid w'_1) + \log P(w_s \mid w'_2) + \cdots + \log P(w_s \mid w'_n)$$

Using Order of Words: RNNs

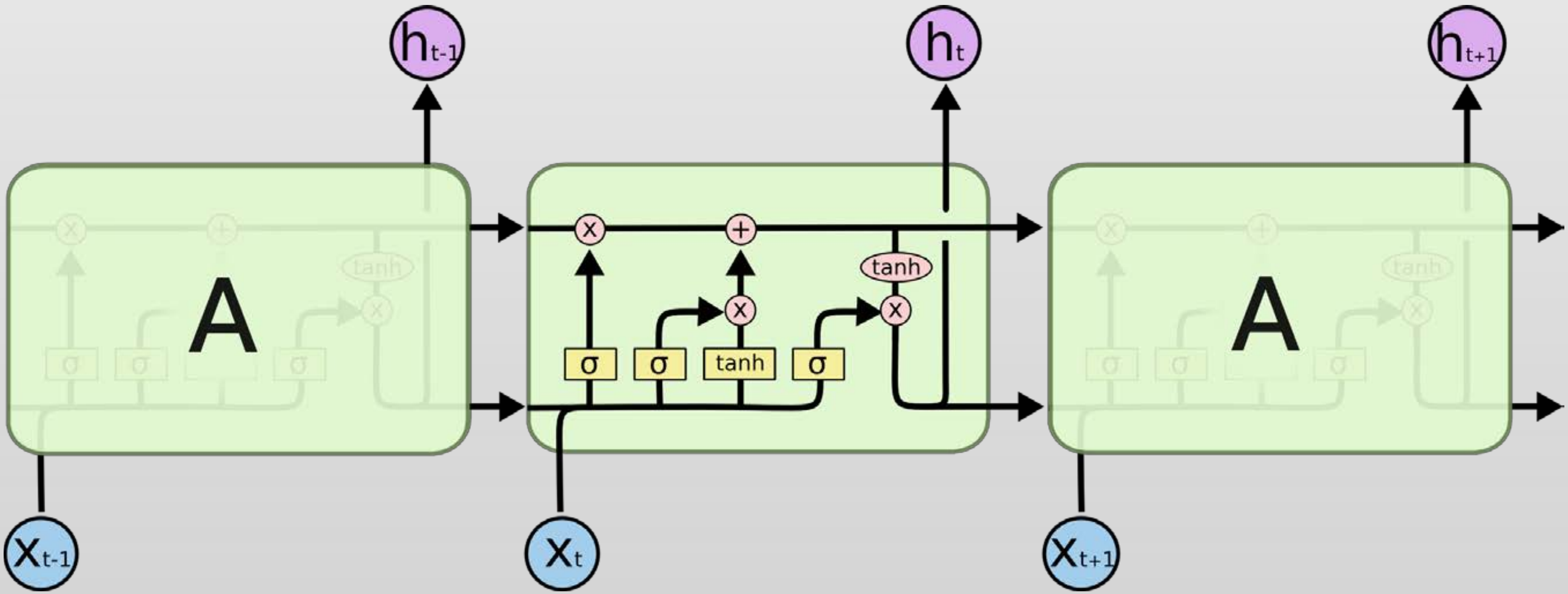
- Can we use the order/sequence of words in the context?
- RNNs!

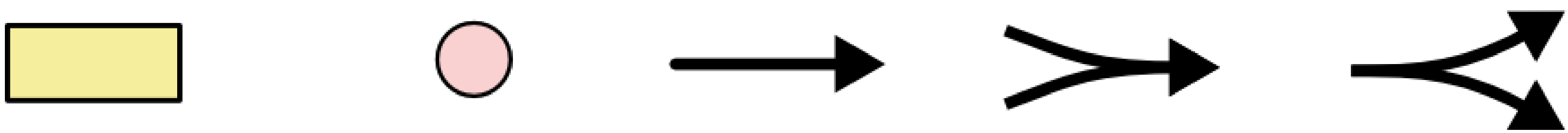


Use long range dependence



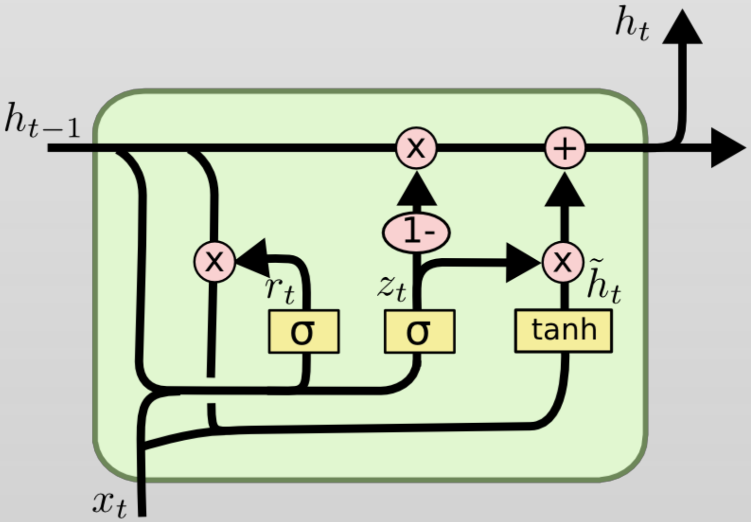
LSTMs!





Neural Network Layer | Pointwise Operation | Vector Transfer | Concatenate | Copy

GRUs



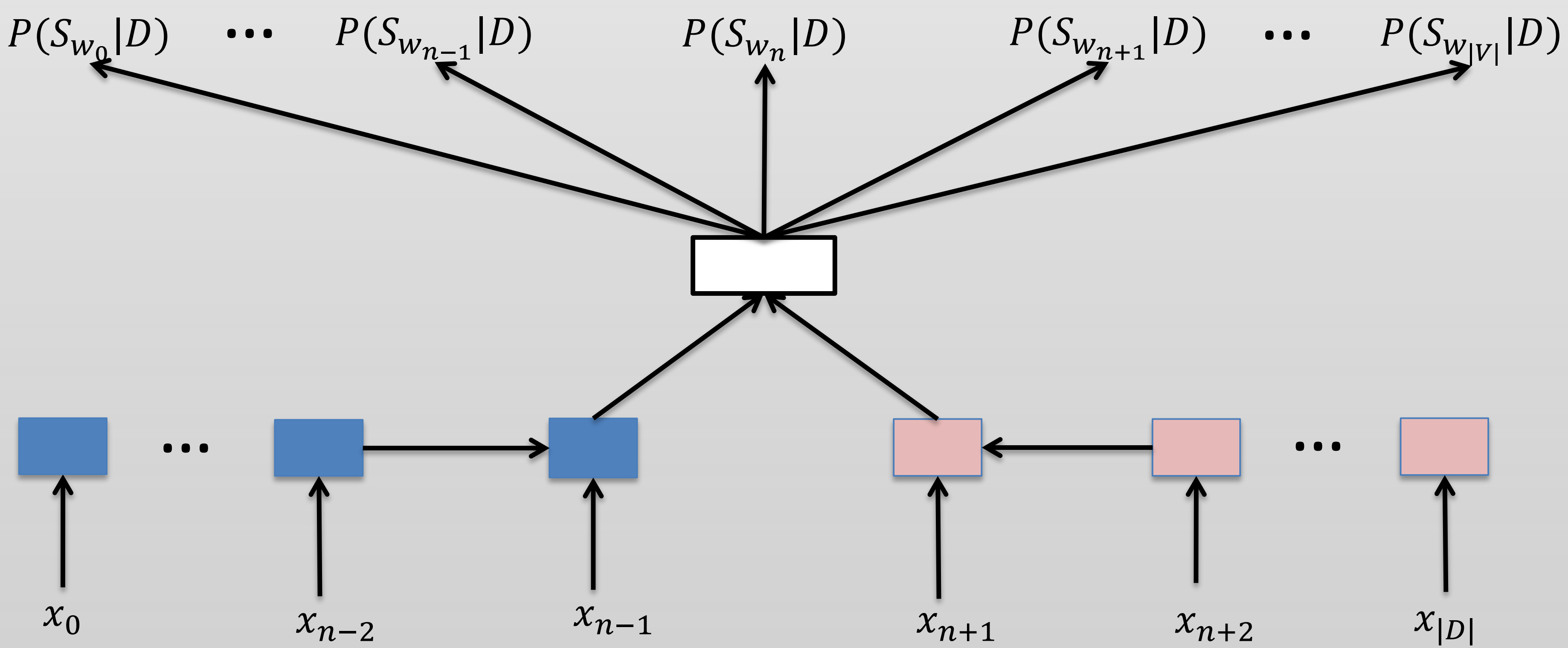
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

WSD: A BLSTM Model



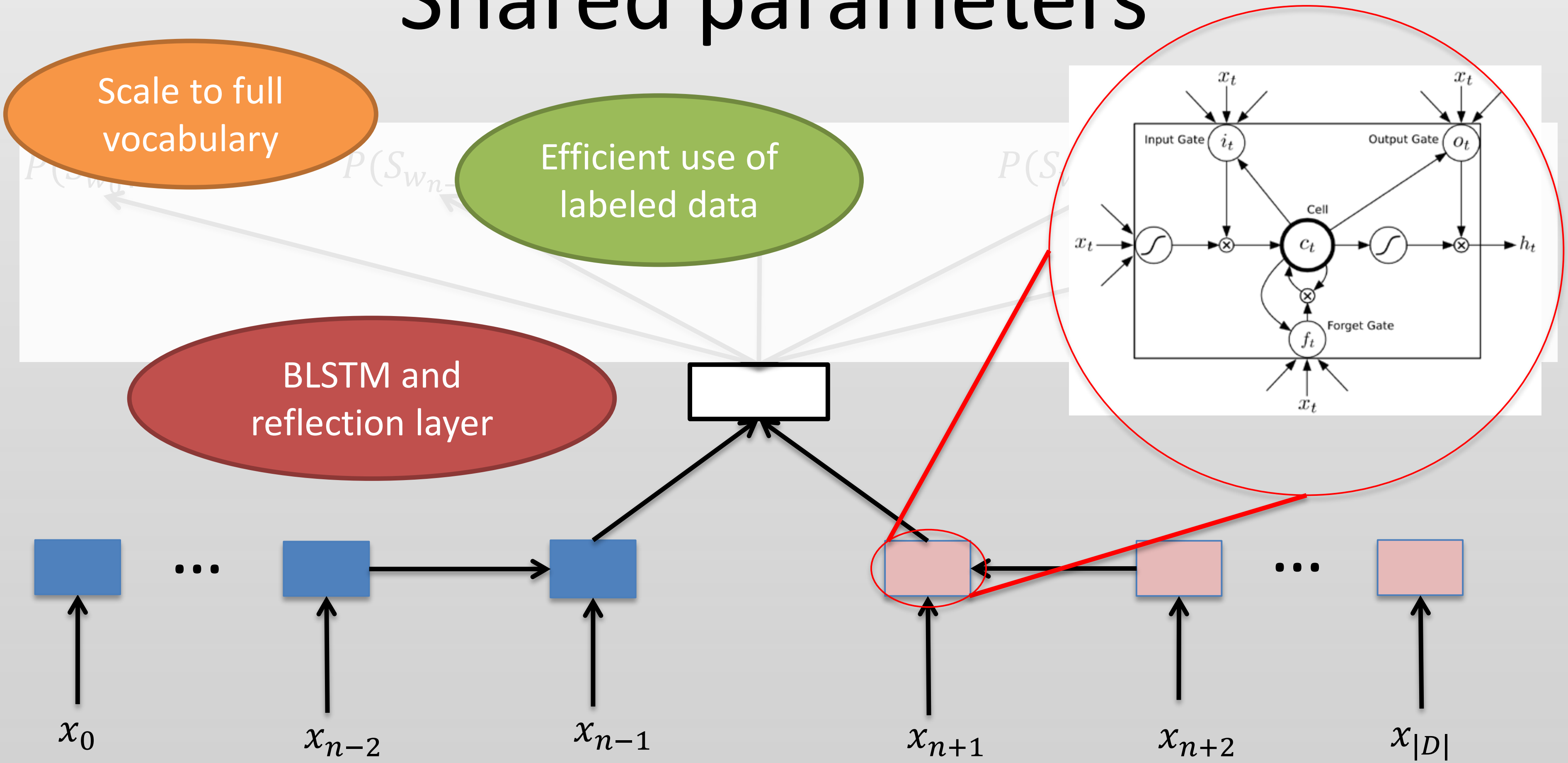
Kågebäck and Salomonsson, *CogAlex, Coling* 2016

https://bitbucket.org/salomons/wsd

S_{w_n}= Senses of word type w_n

x_n= Word token n in document D

Shared parameters



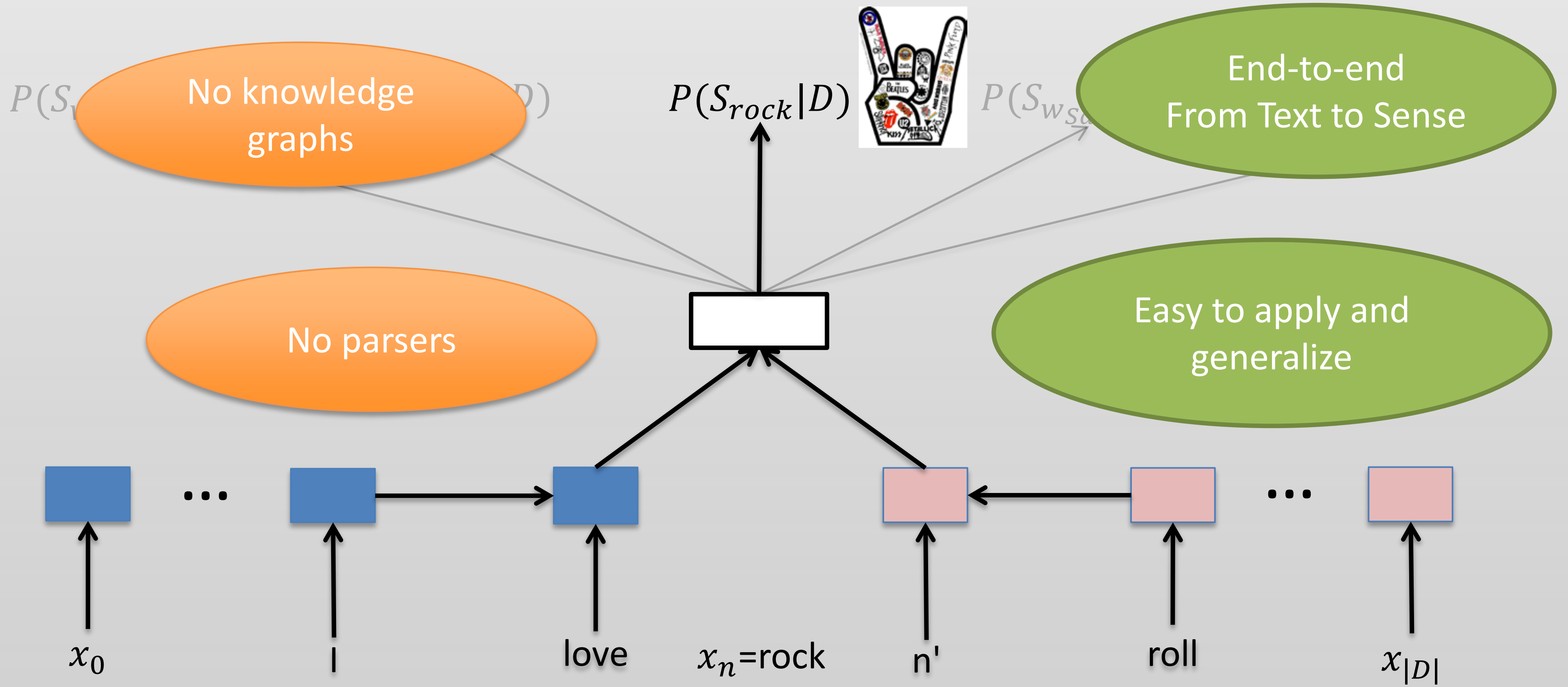
No explicit window

S_{w_n} = Senses of word type w_n

x_n = Word token n in document D

No hand crafted features

Example - rock



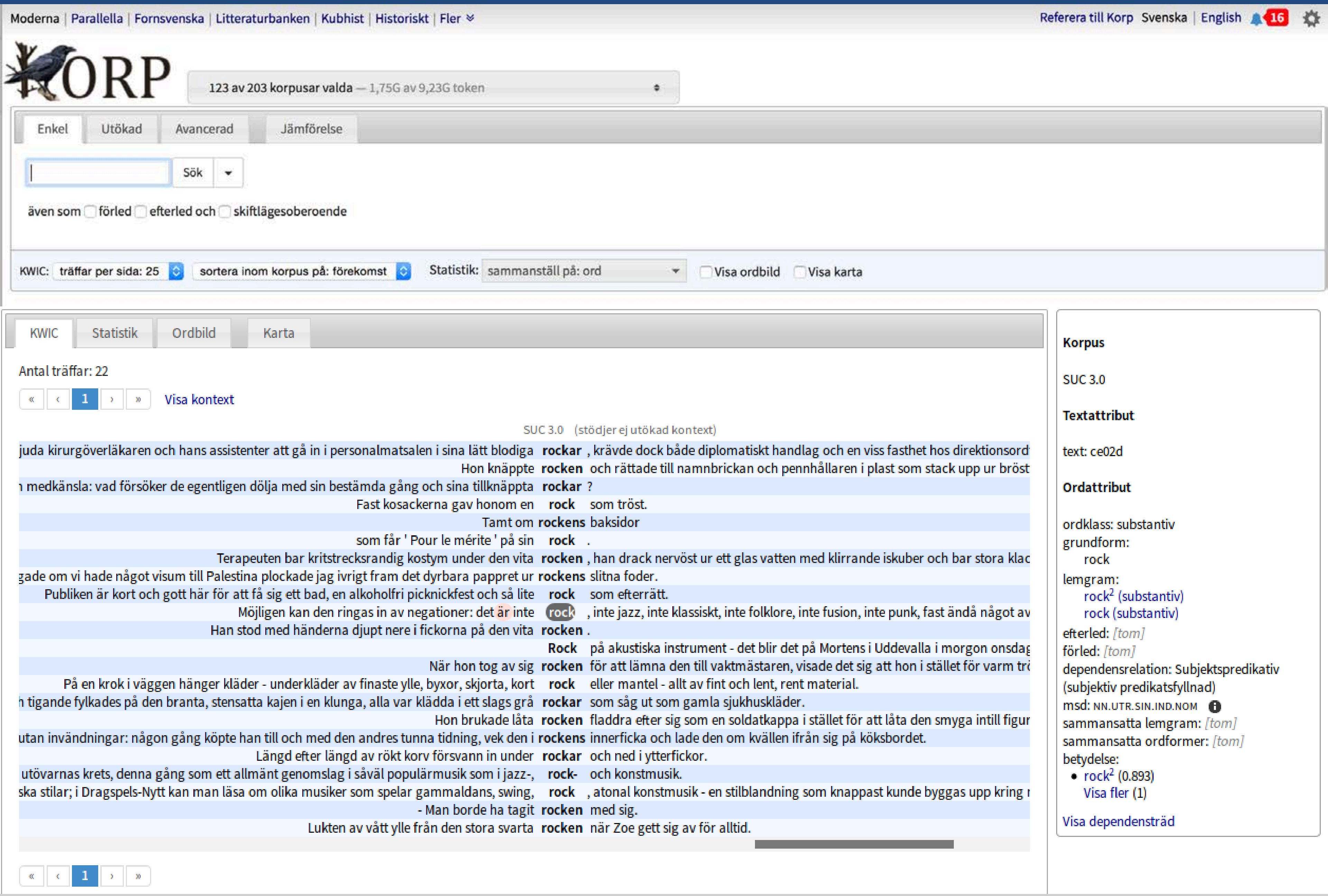
S_{rock}= Senses of word type rock

No part-of-speech

... I love **rock** n' roll ...

Using Sense Vectors

- Sense vectors useful in many downstream NLP tasks, also in machine translation.
- Sense vectors also useful in document summarization: first disambiguate the sense of each word occurrence, then compose sense vectos to form vector for sentences.



Moderna | Parallella | Fornsvenska | Litteraturbanken | Kubhist | Historiskt | Fler
Referera till Korp Svenska | English 16
KORP
123 av 203 korpusar valda — 1,75G av 9,23G token
Enkel Utökad Avancerad Jämförelse
Sök
även som förled efterled och skiftlägesoberoende
KWIC: träffar per sida: 25 sortera inom korpus på: förekomst Statistik: sammanställ på: ord Visa ordbild Visa karta
KWIC Statistik Ordbild Karta
Antal träffar: 22
« ‹ 1 › » Visa kontext
SUC 3.0 (stödjer ej utökad kontext)
juda kirurgöverläkaren och hans assistenter att gå in i personalmatsalen i sina lätt blodiga rockar , krävde dock både diplomatiskt handlag och en viss fasthet hos direktionsord
Hon knäppte rocken och rättade till namnbrickan och pennhållaren i plast som stack upp ur bröst
medkänsla: vad försöker de egentligen dölja med sin bestämda gång och sina tillknäppta rockar ?
Fast kosackerna gav honom en rock som tröst.
Tamt om rockens baksidor
som får ' Pour le mérite ' på sin rock .
Terapeuten bar kritstrecksrandig kostym under den vita rocken , han drack nervöst ur ett glas vatten med klirrande iskuber och bar stora klac
gade om vi hade något visum till Palestina plockade jag ivrigt fram det dyrbara pappret ur rockens slitna foder.
Publiken är kort och gott här för att få sig ett bad, en alkoholfri picknickfest och så lite rock som efterrätt.
Möjligen kan den ringas in av negationer: det är inte rock , inte jazz, inte klassiskt, inte folklore, inte fusion, inte punk, fast ändå något av
Han stod med händerna djupt nere i fickorna på den vita rocken .
Rock på akustiska instrument - det blir det på Mortens i Uddevalla i morgon onsdag
När hon tog av sig rocken för att lämna den till vaktmästaren, visade det sig att hon i stället för varm tr
På en krok i väggen hänger kläder - underkläder av finaste ylle, byxor, skjorta, kort rock eller mantel - allt av fint och lent, rent material.
tigande fylkades på den branta, stensatta kajen i en klunga, alla var klädda i ett slags grå rockar som såg ut som gamla sjukhuskläder.
Hon brukade låta rocken fladdra efter sig som en soldatkappa i stället för att låta den smyga intill figur
utan invändningar: någon gång köpte han till och med den andres tunna tidning, vek den i rockens innerficka och lade den om kvällen ifrån sig på köksbordet.
Längd efter längd av rökt korv försvann in under rockar och ned i ytterfickor.
utövarnas krets, denna gång som ett allmänt genomslag i såväl populärmusik som i jazz-, rock- och konstmusik.
ska stilar; i Dragspels-Nytt kan man läsa om olika musiker som spelar gammaldans, swing, rock , atonal konstmusik - en stilblandning som knappast kunde byggas upp kring
- Man borde ha tagit rocken med sig.
Lukten av vått ylle från den stora svarta rocken när Zoe gett sig av för alltid.
« ‹ 1 › »
Korpus
SUC 3.0
Textattribut
text: ce02d
Ordattribut
ordklass: substantiv
grundform:
rock
lemgram:
rock² (substantiv)
rock (substantiv)
efterled: [tom]
förled: [tom]
dependensrelation: Subjektspredikativ (subjektiv predikatsfyllnad)
msd: NN.UTR.SIN.IND.NOM
sammansatta lemgram: [tom]
sammansatta ordformer: [tom]
betydelse:
• rock² (0.893)
Visa fler (1)
Visa dependensträd

Adaptive Natural Language Processing

WWW

enhance background by focused crawling

Text Data

Find regularities

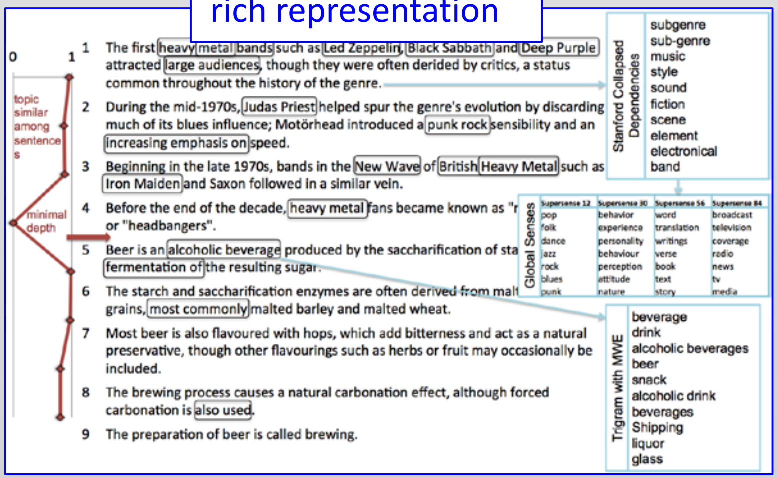
Structure Discovery Algorithms

Annotate regularities in data

Use annotations as features

Adaptive Machine Learning

rich representation

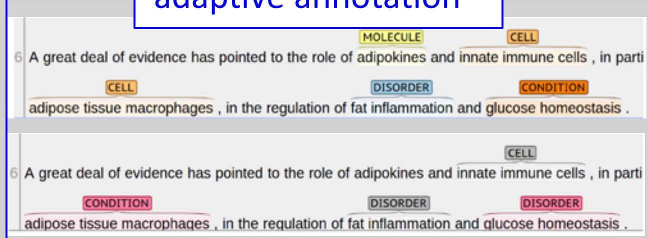


semantic model

jaguar

ISA	**ISA**
vehicle	animal
car	species
brand	wildlife
company	predator
SIM	**SIM**
mercedes	tiger
bmw	leopard
cadillac	lion
jeep	cougar

adaptive annotation



- Adaptive to collection and to human user

Takeaways: Lecture 3

- Word sense induction and disambiguation are fundamental tasks in NLP
- Clustering is a fundamental unsupervised ML technique
- Word, context and sense embeddings are useful tools in these tasks.
- RNN architectures such as LSTMS and GRUs are powerful tools for these tasks

References

- M. Kågebäck et al, **Neural context embeddings for automatic discovery of word senses** *(NAACL 2015 workshop on Vector Space Modeling for NLP)*
- T. Hocking et al, “Clusterpath: an Algorithm for Clustering using Convex Fusion Penalties”, ICML 2011.
- R. Johansson and L. Pena Neto, Embedding a Semantic Network in a Word Space, *NAACL 2015.*
- R. Johansson and L. Pena Neto, Embedding Senses for Efficient Graph Based Word Sense Disambiguation, *TextGraphs@NAACL 2016*
- M. Kågebäck and H. Salomonsson,**Word Sense Disambiguation using a Bidirectional LSTM** *(Coling 2016 Workshop on Cognitive Aspects of the Lexicon (CogALex-V))*