Measuring Distributional Similarity in Context

Georgiana Dinu and Mirella Lapata

Presenter: Dai Quoc Nguyen

Overview

- Introduction
- Vector Space Models
- The Probabilistic Approach
  - Meaning Representation over Latent Senses
  - Latent Dirichlet Allocation
- Experiments and Evaluation
- Conclusion

Overview

- Introduction
- Vector Space Models
- The Probabilistic Approach
  - Meaning Representation over Latent Senses
  - Latent Dirichlet Allocation
- Experiments and Evaluation
- Conclusion

Introduction

- The computation of meaning similarity by using vector space models has found widespread use in many tasks within NLP

- But vector space models do not explicitly identify the different senses (or topics) of words and consequently represent their meaning invariably when they are attested in context
  - *heavy* may refer to an overweight person (e.g., she is heavy)
  - or an excessive user (e.g., some heavy users develop a system)

Introduction

- Model the meaning of words as a distribution over a set of latent senses (or topics)

- Probabilistic framework for representing word meaning and measuring similarity in context

Overview

- Introduction
- Vector Space Models
- The Probabilistic Approach
  - Meaning Representation over Latent Senses
  - Latent Dirichlet Allocation
- Experiments and Evaluation
- Conclusion

Vector Space Models

- The distributional hypothesis
  - words occurring in the same contexts
    → similar meanings

- Distributional semantics
  - building word representations using distributional information in order to capture the meaning

Constructing the models

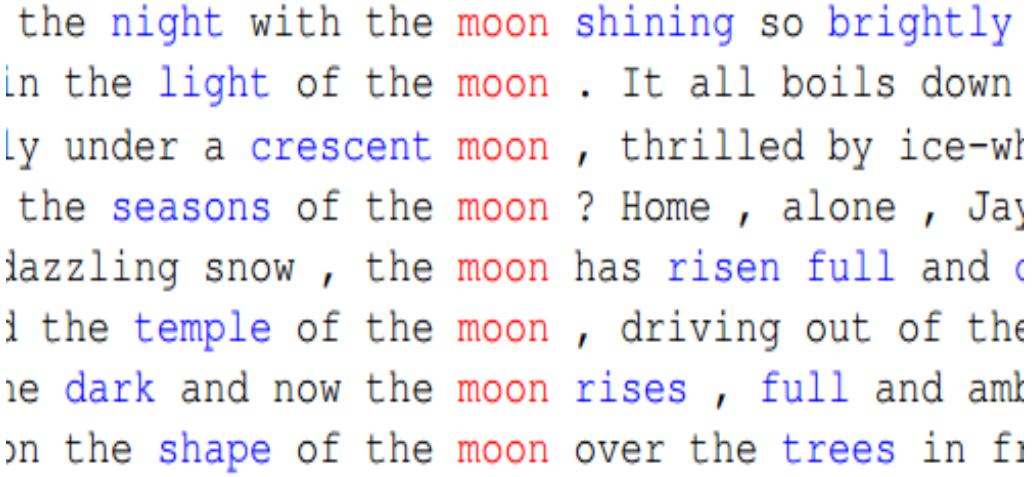
- Co-occurrence counts extraction for target word
- Weighting schemes
- Similarity measures

Extracting co-occurrence counts for target word

- Count how many times each target word occurs in a certain context
  - Contexts are defined by nearby words (in a fixed vocabulary)
  - How often does target word *w* appear near the context *c*
  - For example: *w* appears near *c* if *c* appears within ±5 words of *w*

- Build vectors out of (a function of) these context occurrence counts

	shadow	shine
moon	10	15
sun	8	20



Weighting schemes

- Define how to transform co-occurrence counts of target words *w* and contexts *c* into vector elements of *w*
  - Point-wise Mutual Information (PMI)

    $$PMI(t , c) = \log (P(t , c) / (P(t) * P(c)))$$

  - Term frequency * Inverse document frequency

	shadow	shine
moon	3.01	4.52
sun	2.41	6.02

TF*IDF

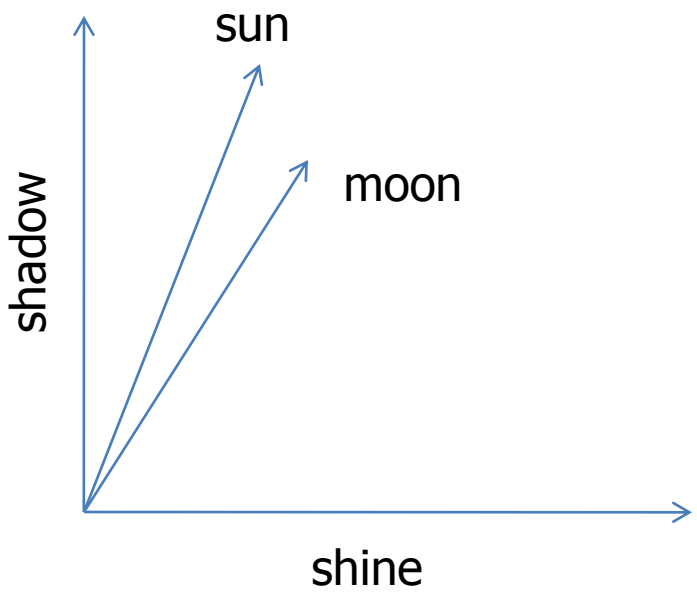
the night with the moon shining so brightly
in the light of the moon . It all boils down
ly under a crescent moon , thrilled by ice-wh
the seasons of the moon ? Home , alone , Jay
dazzling snow , the moon has risen full and c
d the temple of the moon , driving out of the
ne dark and now the moon rises , full and amb
on the shape of the moon over the trees in fr

Similarity measures

- Define how to compute the similarity of word vectors (to approximate similarity in meaning)

  similarity(a, b) = cosine(a, b)

	shadow	shine
moon	3.01	4.52
sun	2.41	6.02



Overview

- Introduction
- Vector Space Models
- The Probabilistic Approach
  - Meaning Representation over Latent Senses
  - Latent Dirichlet Allocation
- Experiments and Evaluation
- Conclusion

Meaning Representation over Latent Senses

- Model the meaning of words as a probability distribution over a set of latent senses (or topics)

- Assume that the target words t_i i : 1...I found in a corpus share a global set of meanings or senses (or topics) $Z = \{z_k/k : 1...K\}$

- Represent a target t_i by the following vector:

$$\mathbf{v}(\mathbf{t_i}) = (\mathbf{P}(\mathbf{z_1}|\mathbf{t_i}), ..., \mathbf{P}(\mathbf{z_K}|\mathbf{t_i}))$$

- Represent the meaning of a target word given a context feature as:

$$\mathbf{v}(\mathbf{t_i}, \mathbf{c_j}) = (\mathbf{P}(\mathbf{z_1}|\mathbf{t_i}, \mathbf{c_j}), ..., \mathbf{P}(\mathbf{z_K}|\mathbf{t_i}, \mathbf{c_j}))$$

Meaning Representation over Latent Senses

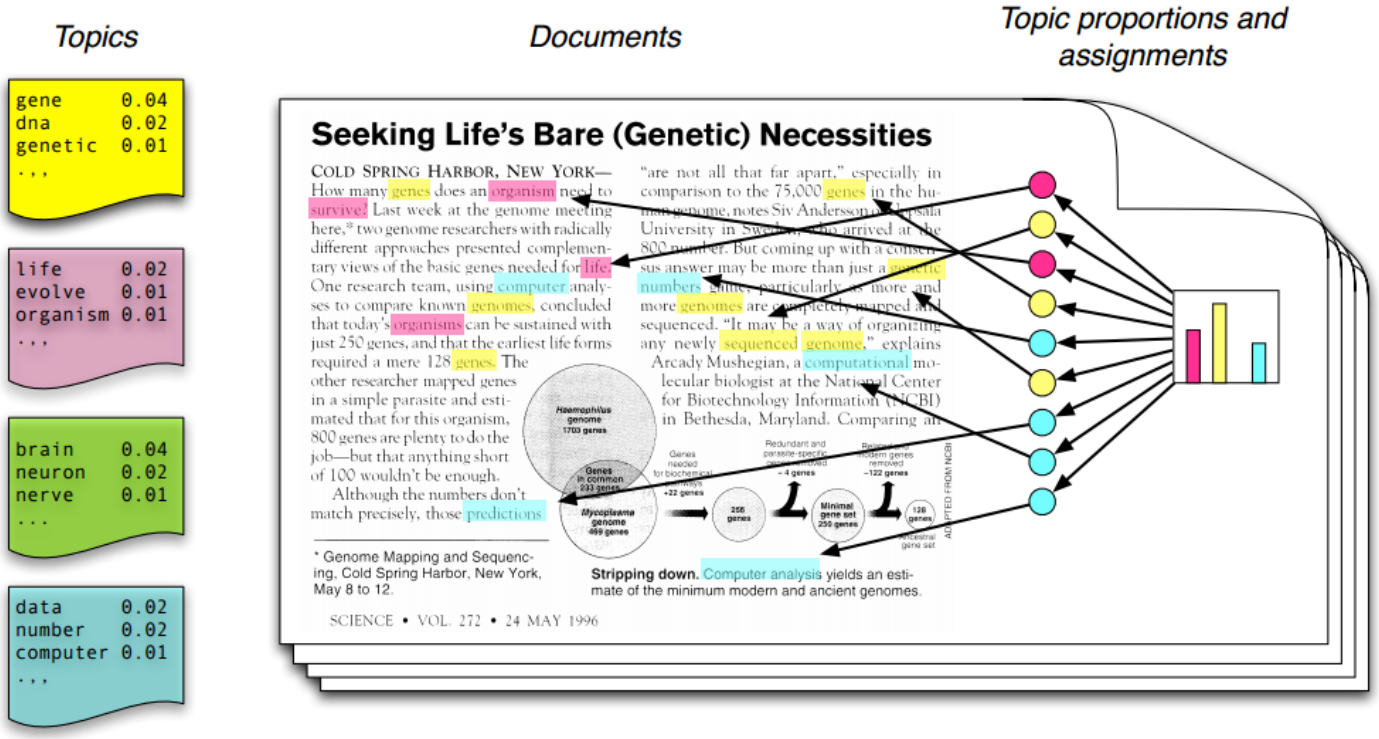
- The conditional probability of context c_j given target t_i and sense z_k:

$$P(z_k|t_i, c_j) = \frac{P(t_i, z_k)P(c_j|z_k, t_i)}{\sum_k P(t_i, z_k)P(c_j|z_k, t_i)}$$

- The term $P(c_j | z_k, t_i)$ is difficult to estimate since it implies learning a total number of K × I J-dimensional distributions
- Assume that target words t_i and context features c_j are conditionally independent given sense z_k:

$$P(z_k|t_i, c_j) \approx \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)}$$

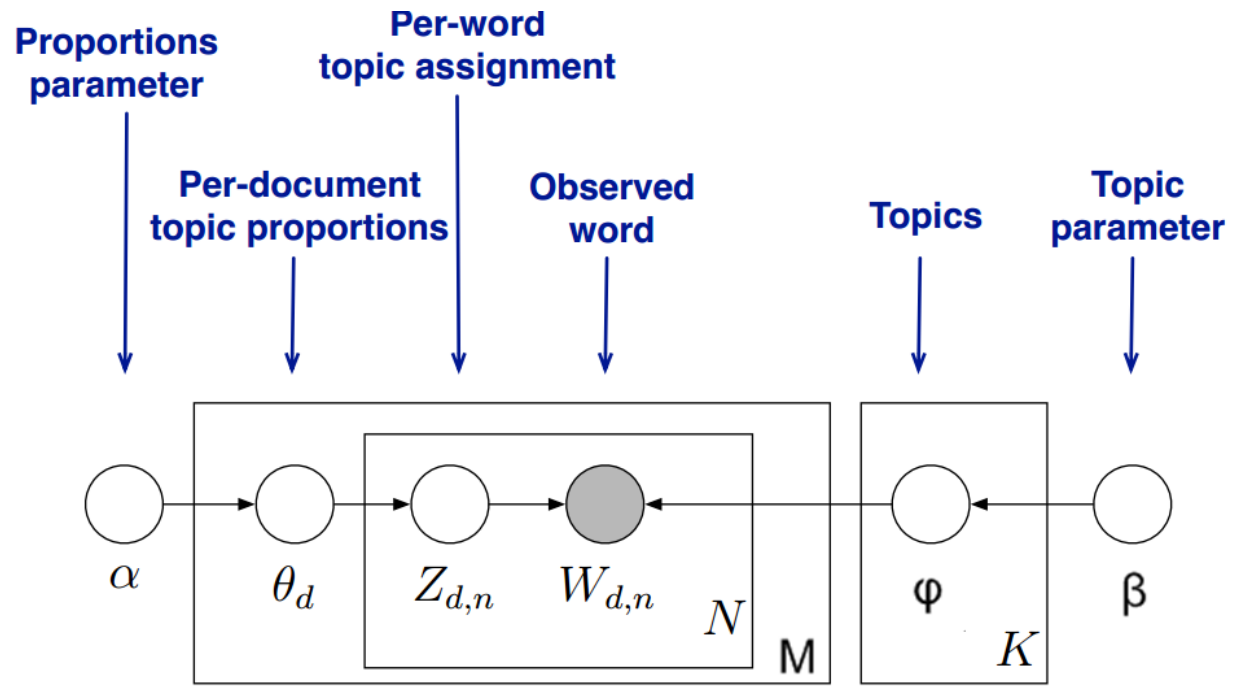
Latent Dirichlet Allocation



- **Intuition**: Documents contain multiple latent topics

- Each document is modeled as a distribution over topics, where each topic is a distribution over words in a fixed vocabulary, and each word is taken from one of those topics

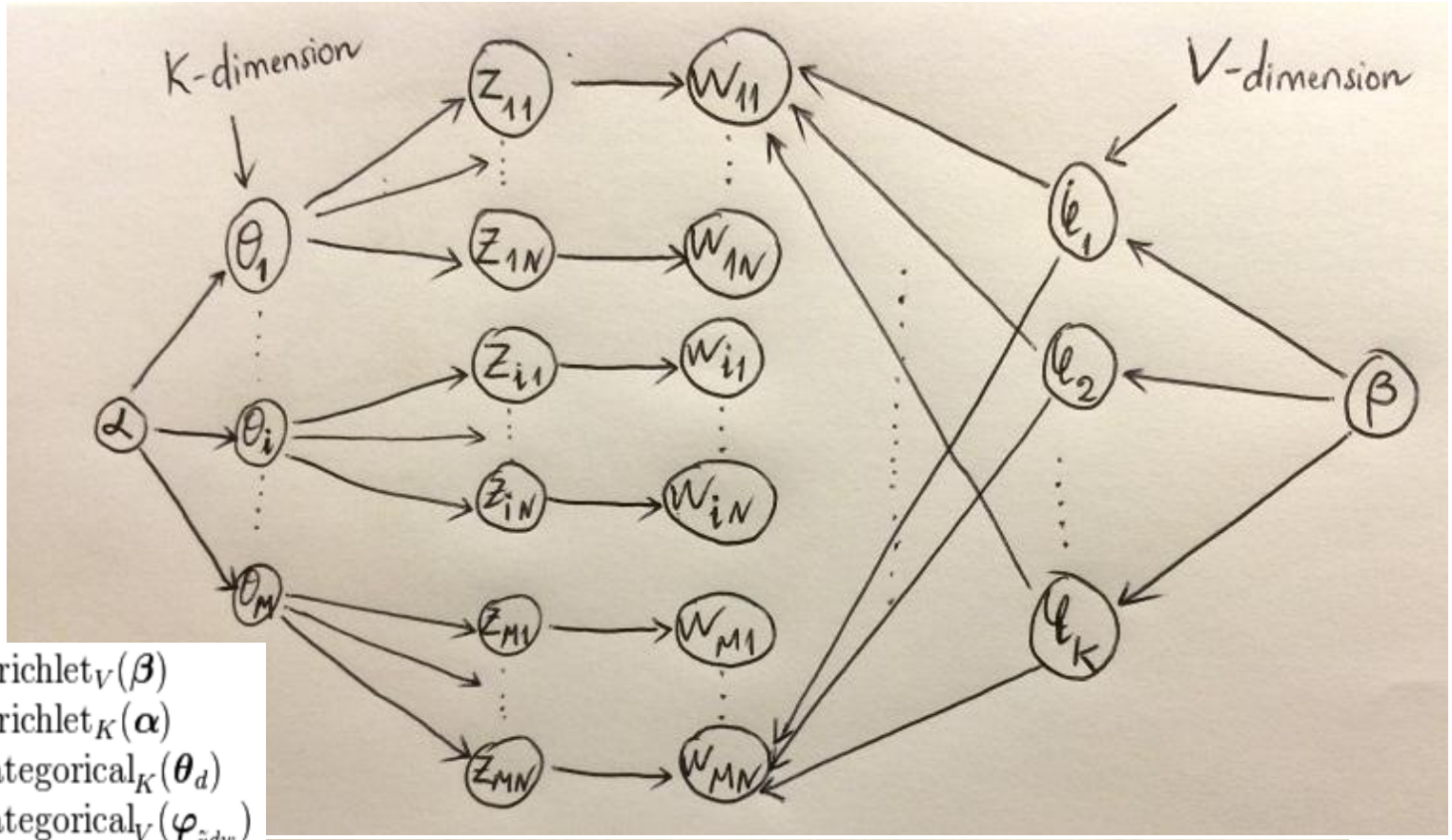
LDA as a graphical model



- The generative process is as follows:

$$
\begin{aligned}
\boldsymbol{\varphi}_{k=1\ldots K} &\sim \text{Dirichlet}_V(\boldsymbol{\beta}) \\
\boldsymbol{\theta}_{d=1\ldots M} &\sim \text{Dirichlet}_K(\boldsymbol{\alpha}) \\
z_{d=1\ldots M, w=1\ldots N_d} &\sim \text{Categorical}_K(\boldsymbol{\theta}_d) \\
w_{d=1\ldots M, w=1\ldots N_d} &\sim \text{Categorical}_V(\boldsymbol{\varphi}_{z_{dw}})
\end{aligned}
$$

LDA as a graphical model



$$\varphi_{k=1\ldots K} \sim \text{Dirichlet}_V(\boldsymbol{\beta})$$
$$\boldsymbol{\theta}_{d=1\ldots M} \sim \text{Dirichlet}_K(\boldsymbol{\alpha})$$
$$z_{d=1\ldots M, w=1\ldots N_d} \sim \text{Categorical}_K(\boldsymbol{\theta}_d)$$
$$w_{d=1\ldots M, w=1\ldots N_d} \sim \text{Categorical}_V(\varphi_{z_{dw}})$$

$$p(W, Z, \theta, \varphi \mid \alpha, \beta) = \prod_{i=1}^{K} p(\varphi_i \mid \beta) \cdot \prod_{d=1}^{M} p(\theta_d \mid \alpha) \cdot \left(\prod_{n=1}^{N} p(z_{dn} \mid \theta_d) \cdot p(w_{dn} \mid z_{dn}, \varphi_{1:K}) \right)$$

LDA as a graphical model

- The posterior distribution:

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

→ intractable in general

→ adopt the Gibbs sampling

- Use LDA to induce senses of target words based on context words

Latent Dirichlet Allocation

- Treat each row t_i in the input matrix transforms as a document
- Train the LDA model on this data to obtain the θ and φ distributions
- θ gives the sense distributions of each target t_i: $\theta_{ik} = P(z_k|t_i)$
- φ *gives* the context-word distribution for each sense z_k: $\varphi_{kj} = P(c_j|z_k)$

$$P(z_k|t_i, c_j) \approx \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)} = \frac{\theta_{ik}\ \varphi_{kj}}{\sum_k \theta_{ik}\ \varphi_{kj}}$$

Overview

- Introduction
- Vector Space Models
- The Probabilistic Approach
  - Meaning Representation over Latent Senses
  - Latent Dirichlet Allocation
- Experiments and Evaluation
- Conclusion

Experiments and Evaluation

- Measure similarity:
  - Scalar product: $\mathrm{sp}(v,w) =< v,w >= \sum_i v_i w_i$
  - Cosine: $\cos(v,w) = \frac{\langle v,w \rangle}{||v||\,||w||}$
  - Inverse Jensen-Shannon (IJS) divergence: $\mathrm{IJS}(v,w) = \frac{1}{\mathrm{JS(v,w)}}$

$$\mathrm{JS(v,w)} = \frac{1}{2}\mathrm{KL}(v|m) + \frac{1}{2}\mathrm{KL}(w|m)$$

$\mathrm{KL}(v|w) = \sum_i v_i log(\frac{v_i}{w_i})$ $\qquad$ $m = \frac{1}{2}(v + w)$

Word Similarity Task

$$sim(t_i, t'_i) = sim(v(t_i), v(t'_i))$$

- Dataset: 353 pairs of words and their similarity scores
- Use Spearman's ρ correlation analysis to examine the relationship between the human ratings and their corresponding vector-based similarity values

Model	Spearman ρ
SVS	38.35
LSA	49.43
NMF	**52.99**
LDA	**53.39**
LSA_{MIX}	49.76
NMF_{MIX}	51.62
LDA_{MIX}	51.97

Lexical Substitution Task

$$sim(t_i, t'_i) = sim(v(t_i, c_j), v(t'_i))$$

- In the lexical substitution task → compute the similarity between t_1 and t_2 given context c : $\{c_1, ..., c_n\}$ as:

$$sim(t_1, t_2|\vec{c}) = sim(v(t_1, c_1), v(t_2)) * ... * sim(v(t_1, c_n), v(t_2))$$

- Dataset: 200 target words, each of which occurs in 10 distinct sentential contexts

Sentences	Substitutes
It is important to apply the herbicide on a *still* day, because spray drift can kill non-target plants.	calm (5) not-windy (1) windless (1)
A movie is a visual document comprised of a series of *still* images.	motionless (3) unmoving (2) fixed (1) stationary (1) static (1)

Lexical Substitution Task

Model	Kendall's τ_b
SVS	11.05
Add-SVS	12.74
Add-NMF	12.85
Add-LDA	12.33
Mult-SVS	14.41
Mult-NMF	13.20
Mult-LDA	12.90
Cont-NMF	14.95
Cont-LDA	13.71
Cont-NMF$_{\text{MIX}}$	**16.01**
Cont-LDA$_{\text{MIX}}$	**15.53**

Overview

- Introduction
- Vector Space Models
- The Probabilistic Approach
  - Meaning Representation over Latent Senses
  - Latent Dirichlet Allocation
- Experiments and Evaluation
- Conclusion

Conclusion

- Presented a general framework for computing similarity in context
- Represented word meaning as a distribution over a set of global senses where contextualized meaning is modeled as a change in this distribution

Thank you for your attention!

Questions?