Embed-Align

Wilker Aziz
(joint work with Miguel Rios)

Universiteit van Amsterdam
w.aziz@uva.nl

April 6, 2017

Advertisement

I am currently working on deep generative models for

- paraphrasing (with Miguel Rios)
- phrase alignment (with Philip Schulz)
- morphology (with Sander de Vroe)
- parsing (with Joost Bastings)
- machine translation (with Amir Kamran)

Advertisement

I am currently working on deep generative models for

- paraphrasing (with Miguel Rios)
- phrase alignment (with Philip Schulz)
- morphology (with Sander de Vroe)
- parsing (with Joost Bastings)
- machine translation (with Amir Kamran)

because

- PGMs make for a very powerful modelling formalism
- neural networks are excellent density estimators
- most of current NLP research relies on complete supervision

Advertisement

I am currently working on deep generative models for

- paraphrasing (with Miguel Rios)
- phrase alignment (with Philip Schulz)
- morphology (with Sander de Vroe)
- parsing (with Joost Bastings)
- machine translation (with Amir Kamran)

because

- PGMs make for a very powerful modelling formalism
- neural networks are excellent density estimators
- most of current NLP research relies on complete supervision

If you want to know more, you will find me at F2.11.

Embed-Align

Motivation

Background

Deep generative models

Lexical paraphrasing

In the event of a chemical spill, 3/4's of the children know that they should ***evacuate*** *as advised on radio, TV, or by people in charge.*

Task: hinges on a form of context-dependent WSD

1. predict candidates to substitute **evacuate**
2. rank them for equivalence

Lexical paraphrasing: one typical solution

Representation learning

1. represent words in context (e.g. embeddings, BiLSTMs, ...)
2. pick your favourite metric m (e.g. cosine)
3. rank vocabulary as a function of $m(\vec{v}(\text{charge}), \vec{v}(w))$

Lexical paraphrasing: one typical solution

Representation learning

1. represent words in context (e.g. embeddings, BiLSTMs, ...)
2. pick your favourite metric m (e.g. cosine)
3. rank vocabulary as a function of $m(\vec{v}(\text{charge}), \vec{v}(w))$

Some known properties

1. seems to recover metrics (enabling word analogies)
   $\vec{v}(\textit{king}) - \vec{v}(\textit{man}) + \vec{v}(\textit{woman}) \approx \vec{v}(\textit{queen})$

Lexical paraphrasing: one typical solution

Representation learning

1. represent words in context (e.g. embeddings, BiLSTMs, ...)
2. pick your favourite metric m (e.g. cosine)
3. rank vocabulary as a function of $m(\vec{v}(\text{charge}), \vec{v}(w))$

Some known properties

1. seems to recover metrics (enabling word analogies)
   $\vec{v}(\textit{king}) - \vec{v}(\textit{man}) + \vec{v}(\textit{woman}) \approx \vec{v}(\textit{queen})$

Some known caveats

1. candidates cannot be **generated** by model
2. disambiguation relies solely on Harris hypothesis
   may not hold for antonyms and closely-related senses

Lexical paraphrasing: another typical solution

Pivoting

1. represent words as distributions over foreign vocabulary
   lexical alignments in two directions
2. candidate set: words sharing common translations
3. ranking: by interpolated scores or a discriminative model

Lexical paraphrasing: another typical solution

Pivoting

1. represent words as distributions over foreign vocabulary
   lexical alignments in two directions
2. candidate set: words sharing common translations
3. ranking: by interpolated scores or a discriminative model

Some known properties

1. strong disambiguation power
   antonyms and closely-related senses are further discriminated

Lexical paraphrasing: another typical solution

Pivoting

1. represent words as distributions over foreign vocabulary
   lexical alignments in two directions
2. candidate set: words sharing common translations
3. ranking: by interpolated scores or a discriminative model

Some known properties

1. strong disambiguation power
   antonyms and closely-related senses are further discriminated

Some known caveats

1. candidates cannot be **generated** by model
2. monolingual context is mostly gone
   no strong results on metric recovery

Lexical paraphrasing: less typical solution

Topic modelling

1. generative model of documents/sentences/words
2. candidate set: can be sampled from the model
3. ranking: by sample frequency or some probabilistic rule

Lexical paraphrasing: less typical solution

Topic modelling

1. generative model of documents/sentences/words
2. candidate set: can be sampled from the model
3. ranking: by sample frequency or some probabilistic rule

Variants

1. hierarchical, nonparametric, multilingual

Lexical paraphrasing: less typical solution

Topic modelling

1. generative model of documents/sentences/words
2. candidate set: can be sampled from the model
3. ranking: by sample frequency or some probabilistic rule

Variants

1. hierarchical, nonparametric, multilingual

Hard to use/induce rich features

What is this talk about?

Deep generative models for lexical paraphrasing

What is this talk about?

Deep generative models for lexical paraphrasing

Word embedding is an unsupervised problem

- we cannot actually observe embeddings
  thus we want to work with generative models

What is this talk about?

Deep generative models for lexical paraphrasing

Word embedding is an unsupervised problem

- we cannot actually observe embeddings
  thus we want to work with generative models

Harris hypothesis seems pretty strong,
but it fails when context is not sufficiently discriminative

- we know where to find additional learning signals

What is this talk about?

Deep generative models for lexical paraphrasing

Word embedding is an unsupervised problem

- we cannot actually observe embeddings
  thus we want to work with generative models

Harris hypothesis seems pretty strong,
but it fails when context is not sufficiently discriminative

- we know where to find additional learning signals

Feature-rich models seem crucial

Embed-Align

Motivation

Background

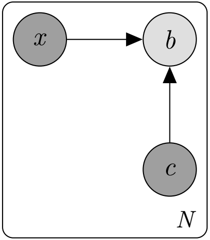
Deep generative models

Supervised embedding models

In the event of a chemical spill, 3/4's of the children know that they should ***evacuate*** *as advised on radio, TV, or by people in charge.*

Supervised embedding models

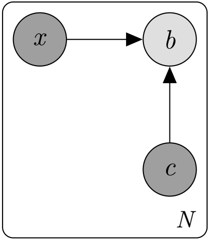
In the event of a chemical spill, 3/4's of the children know that they should ***evacuate*** *as advised on radio, TV, or by people in charge.*



- x is a word
- c a notion of context
- b indicates whether we have observed x and c co-occur

Supervised embedding models

*In the event of a chemical spill, 3/4's of the children know that they should **evacuate** as advised on radio, TV, or by people in charge.*



- x is a word
- c a notion of context
- b indicates whether we have observed x and c co-occur

Artificial strategy to generate labels

- hinges on discriminative power of context
- ad-hoc negative samples
- but highly scalable

Lexical alignment model

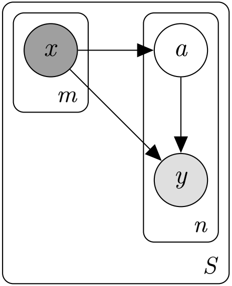
In the event of a chemical spill, 3/4's of the children know that they should ***evacuate*** *as advised on radio, TV, or by people in charge.*

Em caso de vazamento químico, apenas três quartos das crianças estão conscientes de que é necessário evacuar o local, como sugerem rádio, TV, e autoridades.

Lexical alignment model

In the event of a chemical spill, 3/4's of the children know that they should ***evacuate*** *as advised on radio, TV, or by people in charge.*

Em caso de vazamento químico, apenas três quartos das crianças estão conscientes de que é necessário evacuar o local, como sugerem rádio, TV, e autoridades.

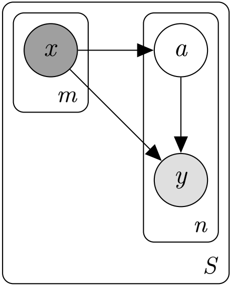


- a selects a position i such that x_i generates y

Lexical alignment model

*In the event of a chemical spill, 3/4's of the children know that they should **evacuate** as advised on radio, TV, or by people in charge.*

Em caso de vazamento químico, apenas três quartos das crianças estão conscientes de que é necessário evacuar o local, como sugerem rádio, TV, e autoridades.

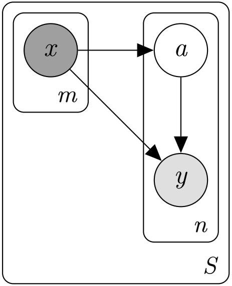


- a selects a position i such that x_i generates y
- lexical alignment provides a proxy to latent "sense labels"

Lexical alignment model

In the event of a chemical spill, 3/4's of the children know that they should ***evacuate*** *as advised on radio, TV, or by people in charge.*

Em caso de vazamento químico, apenas três quartos das crianças estão conscientes de que é necessário evacuar o local, como sugerem rádio, TV, e autoridades.

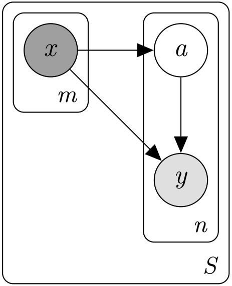


- a selects a position i such that x_i generates y
- lexical alignment provides a proxy to latent "sense labels"
- strong independence assumptions

Lexical alignment model

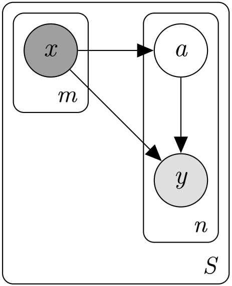
*In the event of a chemical spill, 3/4's of the children know that they should **evacuate** as advised on radio, TV, or by people in charge.*

Em caso de vazamento químico, apenas três quartos das crianças estão conscientes de que é necessário evacuar o local, como sugerem rádio, TV, e autoridades.



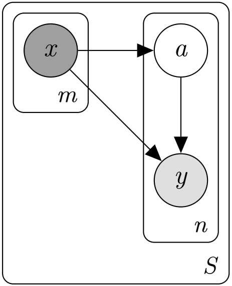
- a selects a position i such that x_i generates y
- lexical alignment provides a proxy to latent "sense labels"
- strong independence assumptions
- no ad-hoc negative labels

Lexical alignment model



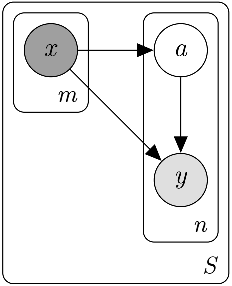
- x_1^m can be represented by our favourite NN architecture

Lexical alignment model



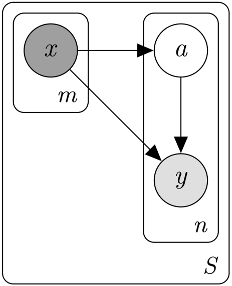
- x_1^m can be represented by our favourite NN architecture
- marginalisation is trivial — though $O(m)$
  $P(Y|x_1^m) = \sum_a P(a|x_1^m)P(Y|x_a)$

Lexical alignment model



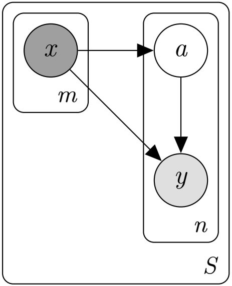
- x_1^m can be represented by our favourite NN architecture
- marginalisation is trivial — though $O(m)$
  $P(Y|x_1^m) = \sum_a P(a|x_1^m)P(Y|x_a)$
- $P(Y|x_a) = \text{softmax}(f(x_a))$ and f is a FFNN

Lexical alignment model



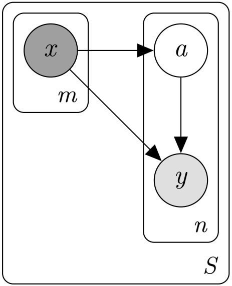
- x_1^m can be represented by our favourite NN architecture
- marginalisation is trivial — though $O(m)$
  $P(Y|x_1^m) = \sum_a P(a|x_1^m)P(Y|x_a)$
- $P(Y|x_a) = \mathrm{softmax}(f(x_a))$ and f is a FFNN
- we take $P(a|x_1^m) = \frac{1}{m}$

Lexical alignment model



- x_1^m can be represented by our favourite NN architecture
- marginalisation is trivial — though $O(m)$
  $P(Y|x_1^m) = \sum_a P(a|x_1^m)P(Y|x_a)$
- $P(Y|x_a) = \text{softmax}(f(x_a))$ and f is a FFNN
- we take $P(a|x_1^m) = \frac{1}{m}$
- maximum likelihood estimate by SGD

Lexical alignment model



- x_1^m can be represented by our favourite NN architecture
- marginalisation is trivial — though $O(m)$
  $P(Y|x_1^m) = \sum_a P(a|x_1^m)P(Y|x_a)$
- $P(Y|x_a) = \mathrm{softmax}(f(x_a))$ and f is a FFNN
- we take $P(a|x_1^m) = \frac{1}{m}$
- maximum likelihood estimate by SGD

But this **does not model** x_1^m

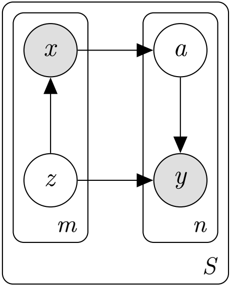
Embed-Align

A generative embed-align model

First, let's generate both sides of the data!

A generative embed-align model

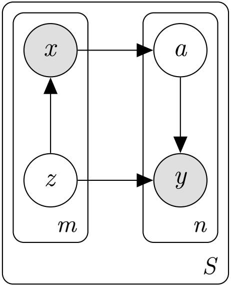
First, let's generate both sides of the data!



- sample latent embeddings $z_1, \ldots, z_m$
  $z_i \sim \mathcal{N}(0, I)$

A generative embed-align model

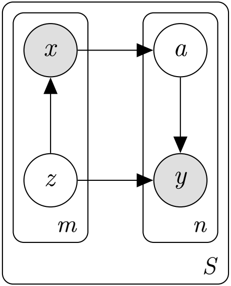
First, let's generate both sides of the data!



- sample latent embeddings $z_1, \ldots, z_m$
  $z_i \sim \mathcal{N}(0, I)$
- generate observation x_i from z_i
  $x_i \sim P(X|Z = z_i)$ (a simple MLP)

A generative embed-align model

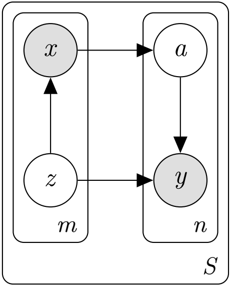
First, let's generate both sides of the data!



- sample latent embeddings $z_1, \ldots, z_m$
  $z_i \sim \mathcal{N}(0, I)$
- generate observation x_i from z_i
  $x_i \sim P(X|Z = z_i)$ (a simple MLP)
- let a select a position in $\{1, \ldots, m\}$
  $a \sim P(A|x_1^m) = \frac{1}{m}$

A generative embed-align model

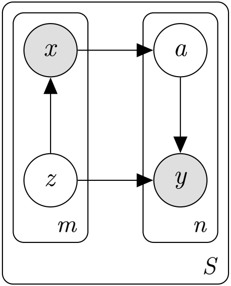
First, let's generate both sides of the data!



- sample latent embeddings $z_1, \ldots, z_m$
  $z_i \sim \mathcal{N}(0, I)$
- generate observation x_i from z_i
  $x_i \sim P(X|Z = z_i)$ (a simple MLP)
- let a select a position in $\{1, \ldots, m\}$
  $a \sim P(A|x_1^m) = \frac{1}{m}$
- and generate y_j from z_a (rather than x_a)
  $y_j \sim P(Y|z_a)$ (a simple MLP)

A generative embed-align model

First, let's generate both sides of the data!

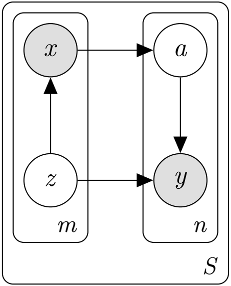


- sample latent embeddings $z_1, \ldots, z_m$
  $z_i \sim \mathcal{N}(0, I)$
- generate observation x_i from z_i
  $x_i \sim P(X|Z = z_i)$ (a simple MLP)
- let a select a position in $\{1, \ldots, m\}$
  $a \sim P(A|x_1^m) = \frac{1}{m}$
- and generate y_j from z_a (rather than x_a)
  $y_j \sim P(Y|z_a)$ (a simple MLP)

Essentially a **variational auto-encoder**

A generative embed-align model

First, let's generate both sides of the data!

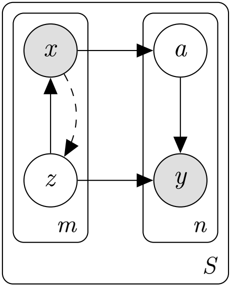


- sample latent embeddings $z_1, \ldots, z_m$
  $z_i \sim \mathcal{N}(0, I)$
- generate observation x_i from z_i
  $x_i \sim P(X|Z = z_i)$ (a simple MLP)
- let a select a position in $\{1, \ldots, m\}$
  $a \sim P(A|x_1^m) = \frac{1}{m}$
- and generate y_j from z_a (rather than x_a)
  $y_j \sim P(Y|z_a)$ (a simple MLP)

Essentially a **variational auto-encoder**
marginalising lexical alignments gathers additional training data

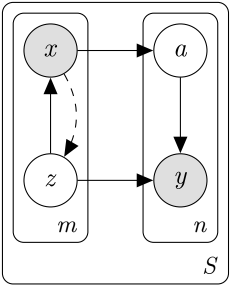
Variational approximation

VI approximates the true posterior $p_\theta(Z|x)$
with an inference model $q_\phi(Z|x)$



Variational approximation

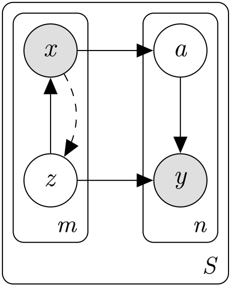
VI approximates the true posterior $p_\theta(Z|x)$
with an inference model $q_\phi(Z|x)$



- Mean field assumption
  $q_\phi(Z_1^m|x_1^m) = \prod_{i=1}^m q_\phi(Z_i|x_1^m)$

Variational approximation

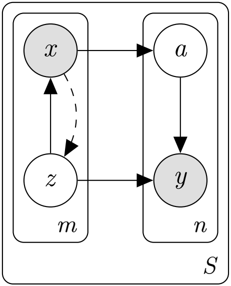
VI approximates the true posterior $p_\theta(Z|x)$
with an inference model $q_\phi(Z|x)$



- Mean field assumption
  $q_\phi(Z_1^m|x_1^m) = \prod_{i=1}^m q_\phi(Z_i|x_1^m)$
- independent local predictions μ_i and σ_i^2
  $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

Variational approximation

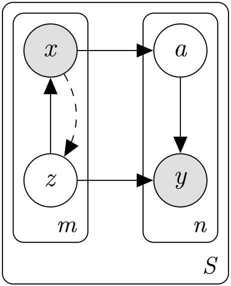
VI approximates the true posterior $p_\theta(Z|x)$
with an inference model $q_\phi(Z|x)$



- Mean field assumption
  $q_\phi(Z_1^m|x_1^m) = \prod_{i=1}^m q_\phi(Z_i|x_1^m)$
- independent local predictions μ_i and σ_i^2
  $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
- $\mu_i = f(r_i(x_1^m))$ and $\sigma_i^2 = g(r_i(x_1^m))$
  - $r_i(x_1^m)$ is the ith BiLSTM representation
  - f and g are FFNNs

Variational approximation

VI approximates the true posterior $p_\theta(Z|x)$
with an inference model $q_\phi(Z|x)$



- Mean field assumption
  $q_\phi(Z_1^m|x_1^m) = \prod_{i=1}^m q_\phi(Z_i|x_1^m)$
- independent local predictions μ_i and σ_i^2
  $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
- $\mu_i = f(r_i(x_1^m))$ and $\sigma_i^2 = g(r_i(x_1^m))$
  - $r_i(x_1^m)$ is the ith BiLSTM representation
  - f and g are FFNNs

Maximise lowerbound on log-likelihood

Preliminary results

English-French parallel data (250,000 sentences)

Model	AER
IBM 1	32.68
Embed-Align	32.06

Table: AER: dx=128, lstm=256, dz=100.

Preliminary results

English-French parallel data (250,000 sentences)

Model	Precision
Discriminative classifier (Giuliano et al, 2007)	69.03
Baseline (Wordnet and corpus frequency)	40.57
Word vecs (Melamud et al, 2015)	27.65
Embed-Align	57.70

Table: LST precision on OOT with constrained candidates

Preliminary results

English-French parallel data (250,000 sentences)

Model	Precision
Discriminative classifier (Giuliano et al, 2007)	6.95
Baseline (Wordnet and corpus frequency)	9.35
Word vecs (Melamud et al, 2015)	8.14
Embed-Align	7.38

Table: LST 1-best precision with constrained candidates

Coming soon

LST with candidates sampled from model

- for some i in x_1^m, repeat for a number of times
  1. $z \sim q(Z_i|x_1^m)$
  2. $x \sim P(X|Z = z)$

Evaluation by word analogy

- hierarchical extension to capture global context

Remarks

I have presented:

- A generative model that "discovers" labelled data

Remarks

I have presented:

- A generative model that "discovers" labelled data
- "Discovery" guided by latent alignments

Remarks

I have presented:

- A generative model that "discovers" labelled data
- "Discovery" guided by latent alignments
- Efficient training (VAE formulation)

Remarks

I have presented:

- A generative model that "discovers" labelled data
- "Discovery" guided by latent alignments
- Efficient training (VAE formulation)

Message I would like to leave:

Architecture design is not the only way to express inductive biases

Remarks

I have presented:

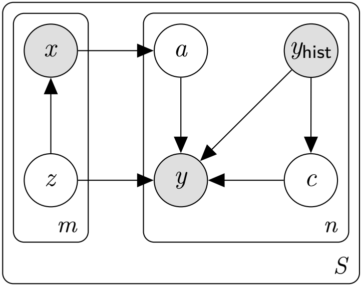
- A generative model that "discovers" labelled data
- "Discovery" guided by latent alignments
- Efficient training (VAE formulation)

Message I would like to leave:

Architecture design is not the only way to express inductive biases

Thanks!

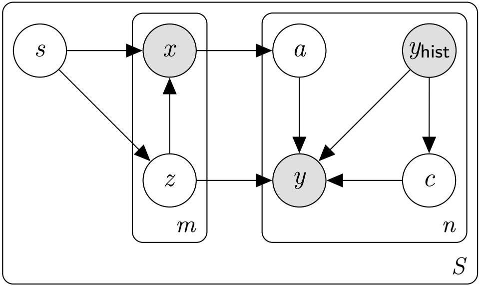
But not every word can be translated



A collocation variable decides between two components

- bilingual component: generates y from z_a
- monolingual component: generates y from monolingual history

But what about Harris hypothesis?

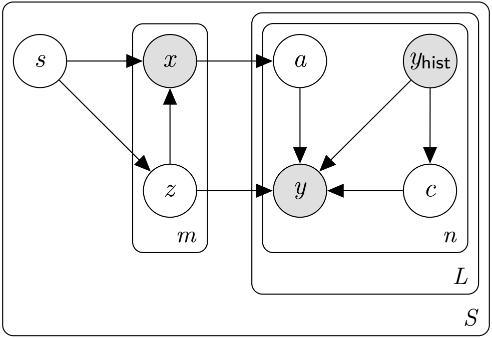


We can capture global context

1. by first sampling $S \sim \mathcal{N}(0, I)$
2. and then $Z \sim \mathcal{N}(\mu(s), \sigma^2(s))$

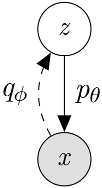
What if bilingual data is not abundant?

We can have multiple languages!



Variational auto-encoders

- Generative model
  $p_\theta(x) = \int p_\theta(z) p_\theta(x|z) \mathrm{d}z$
  where $Z \sim \mathcal{N}(0, I)$
- intractable marginalisation
- variational approximation
  $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x))$



ELBO

$$\log p_\theta(x) \geq -\mathbb{E}_{q_\phi(Z|x)}\left[\log \frac{q_\phi(Z|x)}{p_\theta(Z)}\right] + \mathbb{E}_{q_\phi(Z|x)}\left[\log p_\theta(x|Z)\right]$$

Reparameterised gradient

$$\mathbb{E}_{q_\phi(Z|x)}\left[\log p(x|Z)\right] = \mathbb{E}_{\epsilon \sim N(0,I)}\left[\log p(x|Z = \mu_\phi(x) + \sigma_\phi(x)\epsilon)\right]$$