Deep Generative Model for Joint Alignment and Word Representation

Embedalign

Miguel Rios, Wilker Aziz, Khalil Sima'an
University of Amsterdam
Statistical Language Processing and Learning Lab

June 3, 2018

Outline

Introduction

TL;DR

- Generative model that embeds words in their complete observed context
- Model learns from bilingual sentence-aligned corpora by marginalisation of latent lexical alignments
- Model embeds words as probability densities
- Model shows competitive results on context dependent Natural Language Processing applications

Introduction

Discriminative embedding models

word2vec

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"

Introduction

Discriminative embedding models
word2vec

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"

Fit a binary classifier

- positive examples
- negative examples

Introduction

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

- Limitations

Introduction

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

- Limitations
    - Representation learning is an unsupervised problem we only observe positive/complete context

Introduction

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

- Limitations
  - Representation learning is an unsupervised problem we only observe positive/complete context
  - Distributional hypothesis is strong but fails when context is not discriminative

Introduction

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

- Limitations
  - Representation learning is an unsupervised problem we only observe positive/complete context
  - Distributional hypothesis is strong but fails when context is not discriminative
  - Word senses are collapsed into one vector

Outline

Embedalign

- Generative model to induce word representations

Embedalign

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

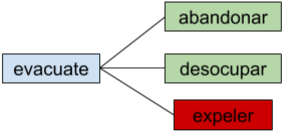
- Generative model to induce word representations
- Learn from positive examples

Embedalign

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

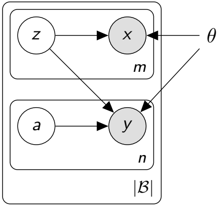
- Generative model to induce word representations
- Learn from positive examples
- Learn from richer (less ambiguous) context
  Foreign text is proxy to sense supervision (Diab and Resnik, 2002)

En caso de un derrame de productos químicos, la mayoría de los niños saben que deben **abandonar** *el lugar según lo aconsejado por las personas a cargo.*



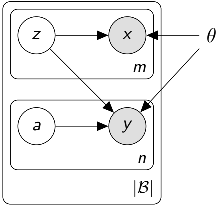
Generative Model

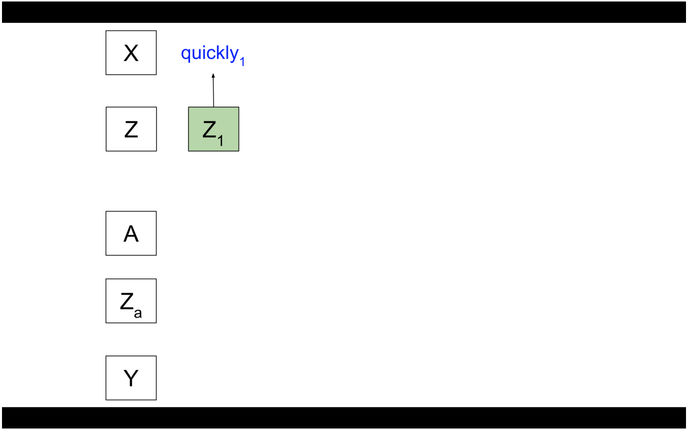
quickly evacuate the area / deje el lugar rápidamente



Generative Model

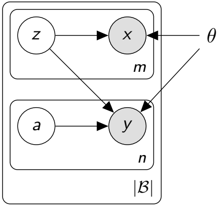
quickly evacuate the area / deje el lugar rápidamente

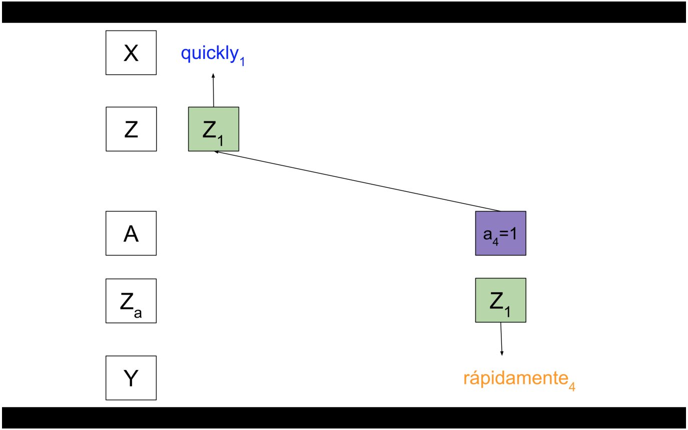




Generative Model

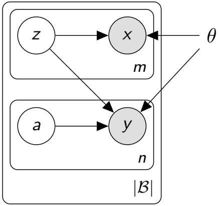
quickly evacuate the area / deje el lugar rápidamente

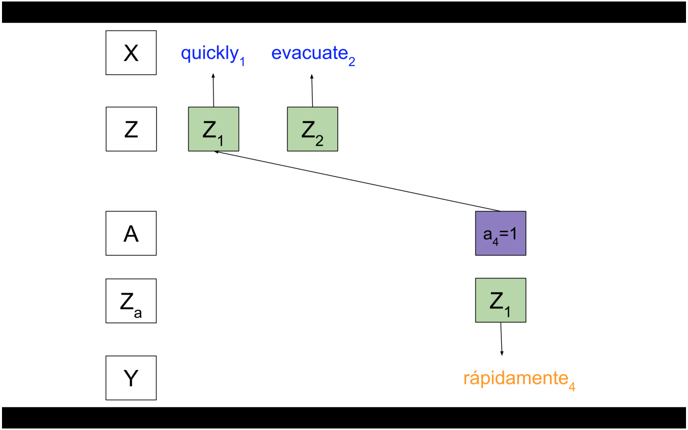




Generative Model

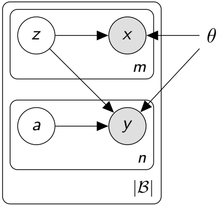
quickly evacuate the area / deje el lugar rápidamente

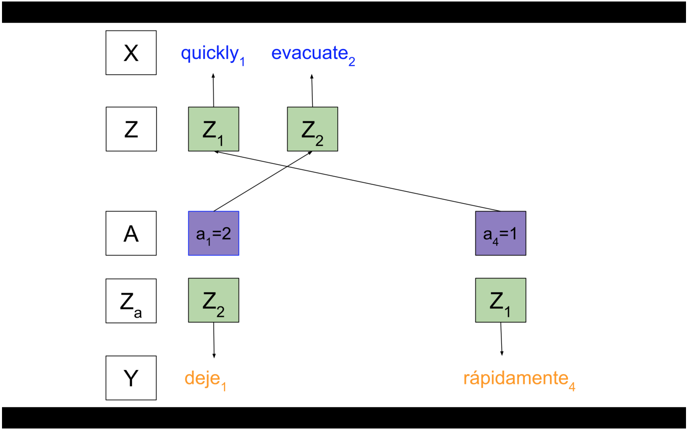




Generative Model

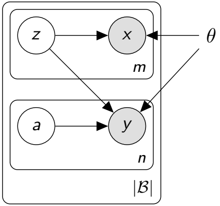
quickly evacuate the area / deje el lugar rápidamente

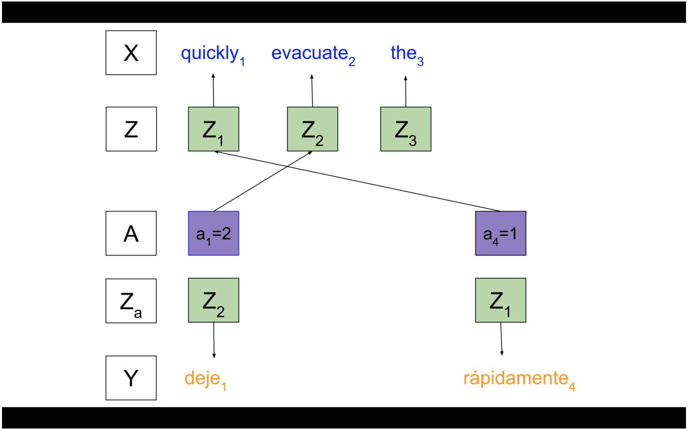




Generative Model

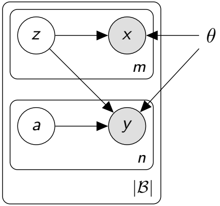
quickly evacuate the area / deje el lugar rápidamente

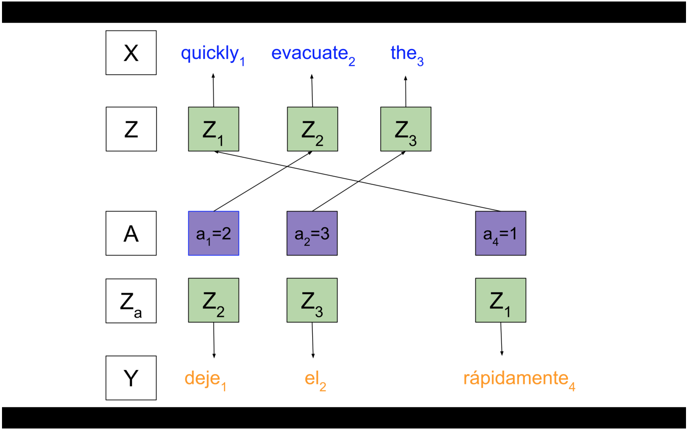




Generative Model

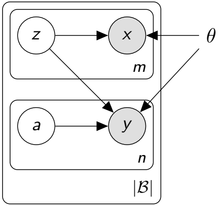
quickly evacuate the area / deje el lugar rápidamente

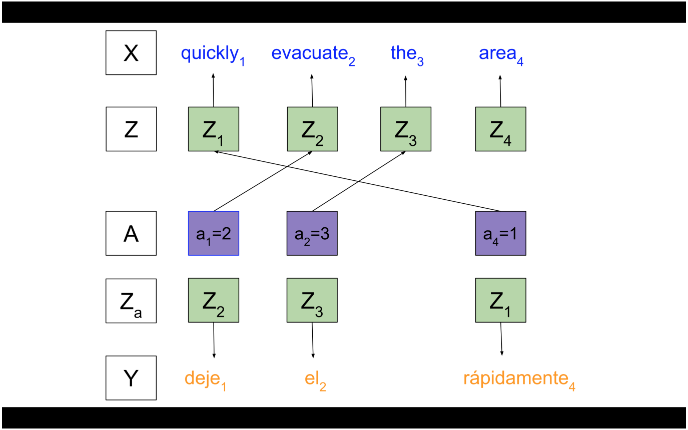




Generative Model

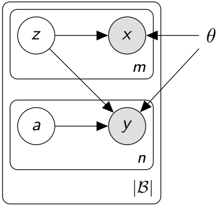
quickly evacuate the area / deje el lugar rápidamente

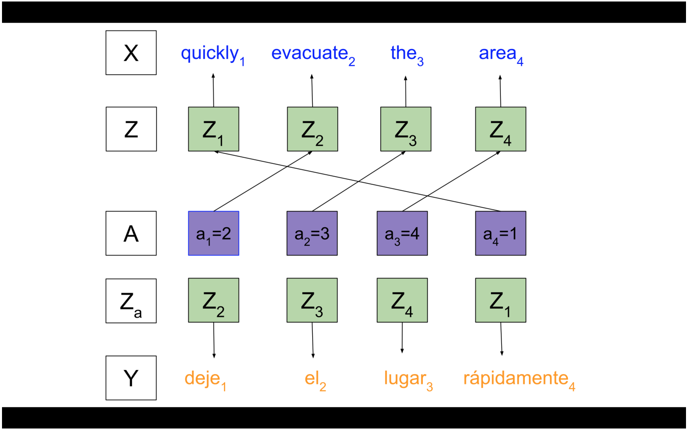




Generative Model

quickly evacuate the area / deje el lugar rápidamente



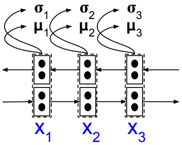


Learning

1. Read sentence

Learning

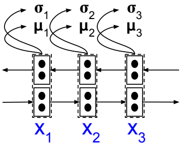
1. Read sentence
2. Predict posterior mean μ_i and std σ_i



evacuate$_1$ the$_2$ area$_3$

Learning

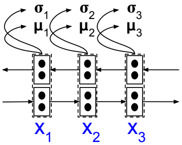
1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$



evacuate$_1$ the$_2$ area$_3$

Learning

1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

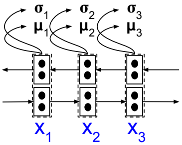


evacuate$_1$ the$_2$ area$_3$

4. Predict categorical distributions

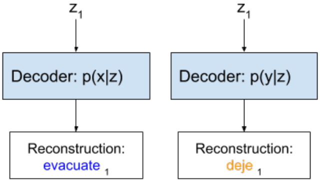
Learning

1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$



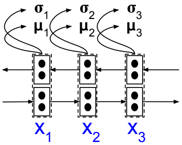
evacuate$_1$ the$_2$ area$_3$

4. Predict categorical distributions
5. Generate observations
   evacuate$_1$ the$_2$ area$_3$ / deje$_1$ el$_2$ lugar$_3$



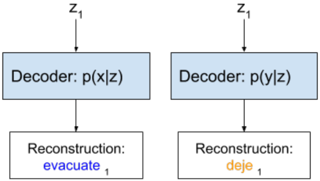
Learning

1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

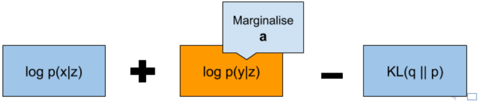


evacuate$_1$ the$_2$ area$_3$

4. Predict categorical distributions
5. Generate observations
   evacuate$_1$ the$_2$ area$_3$ / deje$_1$ el$_2$ lugar$_3$



6. Maximise a lowerbound on likelihood
   (Kingma and Welling, 2014)

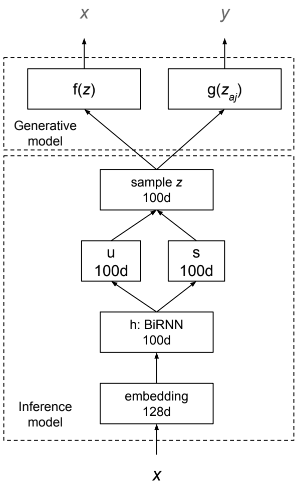


Outline

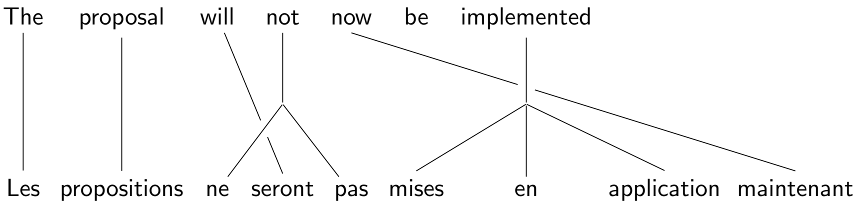
Data

Corpus	Sentence pairs (million)	Tokens (million)
Europarl EN-FR	1.7	42.5
Europarl EN-DE	1.7	43.5

Architecture



Word Alignment



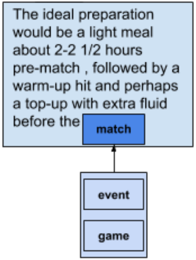
Word Alignment

- Model selection on Dev set

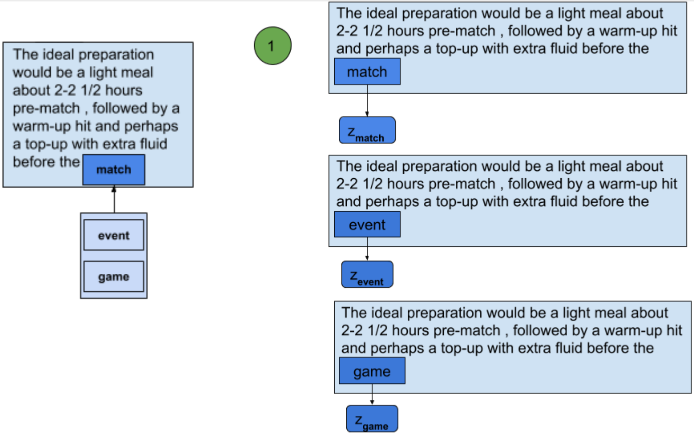
AER ↓

Model	En-Fr	En-De
IBM1	32.45	46.71
IBM2	22.61	40.11
EmbAlign	29.43 ± 1.84	48.09 ± 2.12

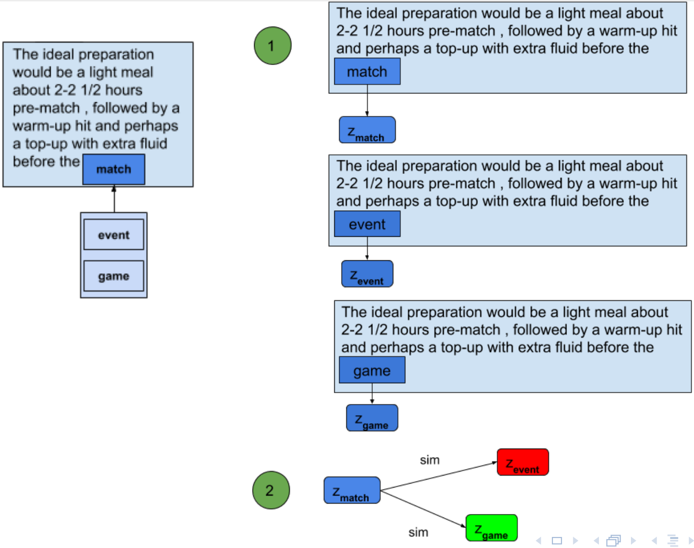
Lexical Substitution



Lexical Substitution



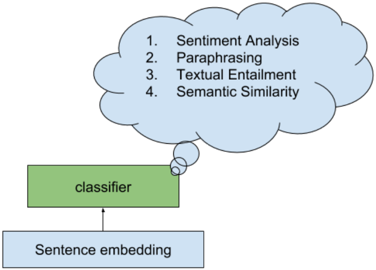
Lexical Substitution



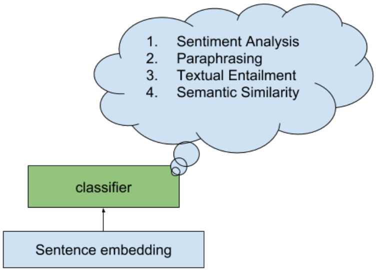
Lexical Substitution

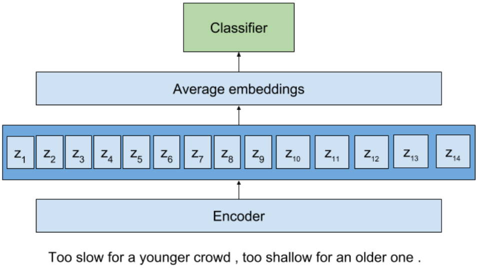
Model	GAP ↑	Training size
Random	30.0	
SkipGram (Melamud et al., 2015)	44.9	ukWaC-2B
BSG (Bražinskas et al., 2017)	46.1	ukWaC-2B
En	21.31 ± 1.05	
En-Fr	42.19 ± 0.57	Euro-42M
En-De	42.07 ± 0.47	

Sentence Evaluation (SentEval)

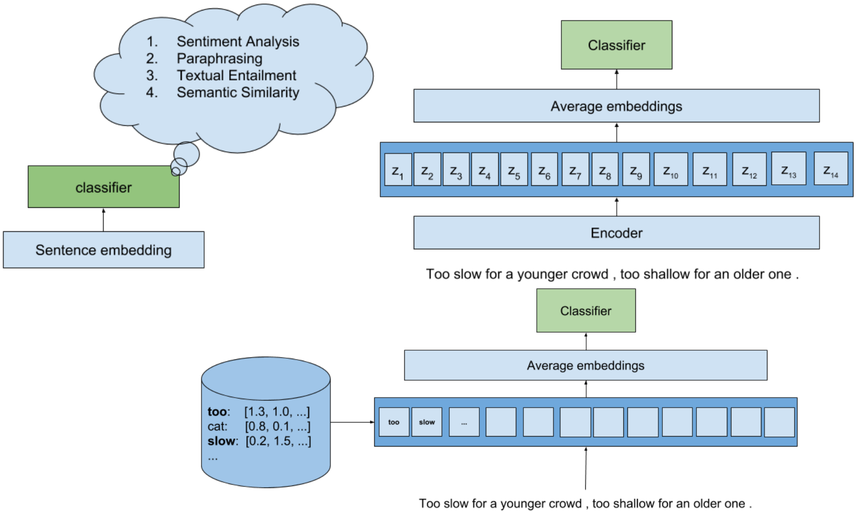


Sentence Evaluation (SentEval)





Sentence Evaluation (SentEval)



Sentence Evaluation (SentEval)

						ACC ↑	ACC/F1 ↑	CORR ↑		CORR ↑
Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STS14
SkipGram$_{\text{En}}$	**70.96**	**76.16**	**87.24**	**86.87**	**73.64**	65.20	70.7/80.1	0.710	76.2	0.45/0.49
En	57.5	67.1	72.0	70.8	57.0	58.0	70.6/80.3	0.648	74.4	0.59/0.59
En-Fr	64.0	71.8	79.1	81.5	64.7	58.4	72.1/81.2	0.682	74.6	0.60/0.59
En-De	62.6	68.0	77.3	82.0	65.0	66.8	70.4/79.8	0.681	75.5	0.58/0.58
Combo	66.1	72.4	82.4	84.4	69.8	**69.0**	**71.9**/**80.6**	**0.727**	**76.3**	**0.62**/**0.61**

Sentence Evaluation (SentEval)

						ACC ↑	ACC/F1 ↑	CORR ↑		CORR ↑
Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STS14
SkipGram$_{\text{En}}$	**70.96**	**76.16**	**87.24**	**86.87**	**73.64**	65.20	70.7/80.1	0.710	76.2	0.45/0.49
En	57.5	67.1	72.0	70.8	57.0	58.0	70.6/80.3	0.648	74.4	0.59/0.59
En-Fr	64.0	71.8	79.1	81.5	64.7	58.4	72.1/81.2	0.682	74.6	0.60/0.59
En-De	62.6	68.0	77.3	82.0	65.0	66.8	70.4/79.8	0.681	75.5	0.58/0.58
Combo	66.1	72.4	82.4	84.4	69.8	**69.0**	**71.9**/**80.6**	**0.727**	**76.3**	**0.62**/**0.61**
SkipGram (Conneau et al., 2017)	**77.7**	**79.8**	**90.9**	**88.3**	**79.7**	**83.6**	**72.5**/**81.4**	**0.803**	**78.7**	**0.65**/**0.64**
NMT$_{\text{En-Fr}}$ (Conneau et al., 2017)	64.7	70.1	84.8	81.5	-	82.8	-	-	-	0.42/0.43

Outline

Conclusions

- Generative training

Conclusions

- Generative training
  - model learns form positive examples

Conclusions

- Generative training
  - model learns form positive examples
  - no need for context window

Conclusions

- Generative training
  - model learns form positive examples
  - no need for context window
- Translation data

Conclusions

- Generative training
  - model learns form positive examples
  - no need for context window
- Translation data
  - less ambiguous embeddings

Conclusions

- Generative training
  - model learns form positive examples
  - no need for context window
- Translation data
  - less ambiguous embeddings
  - model helps with semantic tasks e.g. paraphrasing

Future Work

- We modify alignment distribution

Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER

Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER
- We model word and sentence embeddings

Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER
- We model word and sentence embeddings
  - **M**ovie **R**eviews 66.10 $\rightarrow$ 70.55 ACC

Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 → 18.20 AER
- We model word and sentence embeddings
  - **M**ovie **R**eviews 66.10 → 70.55 ACC
  - **M**icrosoft **P**araphrase 71.90/80.6 → 72.93/81.27 ACC/F1

Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 → 18.20 AER
- We model word and sentence embeddings
  - **M**ovie **R**eviews 66.10 → 70.55 ACC
  - **M**icrosoft **P**araphrase 71.90/80.6 → 72.93/81.27 ACC/F1
  - **S**ick **R** 0.727 → 0.770 CORR

Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 → 18.20 AER
- We model word and sentence embeddings
  - **M**ovie **R**eviews 66.10 → 70.55 ACC
  - **M**icrosoft **P**araphrase 71.90/80.6 → 72.93/81.27 ACC/F1
  - **S**ick **R** 0.727 → 0.770 CORR
- We **will** expand the distributional context to multiple foreign languages at once

DGM4NLP research at UvA-SLPL

- Try pre-trained Europarl model on SentEval: `https://github.com/uva-slpl/embedalign/blob/master/notebooks/senteval_embedalign.ipynb`

DGM4NLP research at UvA-SLPL

- Try pre-trained Europarl model on SentEval: `https://github.com/uva-slpl/embedalign/blob/master/notebooks/senteval_embedalign.ipynb`
- ACL-18 tutorial Variational Inference and Deep Generative Models: `http://acl2018.org/tutorials/`

Lexical Substitution Complete

	GAP ↑		
Model	cos	KL	Training size
RANDOM	30.0	-	
SKIPGRAM (Melamud et al., 2015)	44.9	-	ukWaC-2B
BSG (Bražinskas et al., 2017)	-	46.1	ukWaC-2B
EN	21.31 ± 1.05	27.64 ± 0.40	
EN-FR	42.19 ± 0.57	41.61 ± 0.55	Euro-2M
EN-DE	42.07 ± 0.47	40.93 ± 0.59	