comparing word embedding models

elena fano, jussi karlgren, joakim nivre

uppsala university, gavagai, kth

eRisk, september 2019

compare three word representations:
random indexing, GloVe, and ELMo

compare three word representations:
random indexing, GloVe, and ELMo

and two categorisers:
using linear regression and multi-layer perceptrons

compare three word representations:
random indexing, GloVe, and ELMo

and two categorisers:
using linear regression and multi-layer perceptrons

for
early identification of eating disorder in internet forum post authors

challenges:

- vocabulary variation
- false positives
- most authors discuss many topics
- early detection
- use case vs evaluation metric validity

challenges:

- vocabulary variation
- false positives
- most authors discuss many topics
- early detection
- use case vs evaluation metric validity

challenges:

- vocabulary variation
- false positives
- most authors discuss many topics
- early detection
- use case vs evaluation metric validity

challenges:

- vocabulary variation
- false positives
- most authors discuss many topics
- early detection
- use case vs evaluation metric validity

challenges:

- vocabulary variation
- false positives
- most authors discuss many topics
- early detection
- use case vs evaluation metric validity

challenges:

- vocabulary variation
- false positives
- most authors discuss many topics
- early detection
- use case vs evaluation metric validity

We have two experimental foci:

We have two experimental foci: (1) the representation of lexical items,

We have two experimental foci: (1) the representation of lexical items,and (2) text classification given such representations.

why represent lexical items as vectors?

- generalisation from strings to concepts → recall enhancement
- easily learnable and handy representation it's a trap!
- can capture synonymy and association but not at the same time!
- choice of model should not matter: similar data
  - context and distributiona definition
  - learning curves

why represent lexical items as vectors?

- generalisation from strings to concepts → recall enhancement
- easily learnable and handy representation it's a trap!
- can capture synonymy and association but not at the same time!
- choice of model should not matter: similar data
  - context and distributiona definition
  - learning curves

why represent lexical items as vectors?

- generalisation from strings to concepts → recall enhancement
- easily learnable and handy representation it's a trap!
- can capture synonymy and association but not at the same time!
- choice of model should not matter: similar data
  - context and distributiona definition
  - learning curves

why represent lexical items as vectors?

- generalisation from strings to concepts → recall enhancement
- easily learnable and handy representation it's a trap!
- can capture synonymy and association but not at the same time!
- choice of model should not matter: similar data
  - context and distributiona definition
  - learning curves

why represent lexical items as vectors?

- generalisation from strings to concepts → recall enhancement
- easily learnable and handy representation it's a trap!
- can capture synonymy and association but not at the same time!
- choice of model should not matter: similar data
  - context and distributiona definition
  - learning curves

why represent lexical items as vectors?

- generalisation from strings to concepts → recall enhancement
- easily learnable and handy representation it's a trap!
- can capture synonymy and association but not at the same time!
- choice of model should not matter: similar data
  - context and distributiona definition
  - learning curves

baseline representation

randomly initialized meaningless 100-dimensional word vectors (from Keras)

random indexing

sparse distributed memory model

neurophysiologically plausible, efficiently implementable, online learning, non-compiling, fixed-dimensional

2000-dimensional, trained on social and editorial media

2 + 2 context, OOV terms have empty vectors

- Pentti Kanerva. Sparse distributed memory. MIT press, 1988.
- Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. CogSci.
- Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a Means to Encode Order in Word Space. CogSci.

GloVe

global vectors

pretrained precursor to transfer learning, proven useful in many academic experiments

200-dimensional, pretrained on microblog posts by Stanford NLP group

15-word window, OOV terms with dummy vectors with average distribution

---

- Jeffrey Pennington, Richard Socher and Christopher Manning. 2014. Glove: Global vectors for word representation. EMNLP

ELMo

embeddings from language models

transfer learning

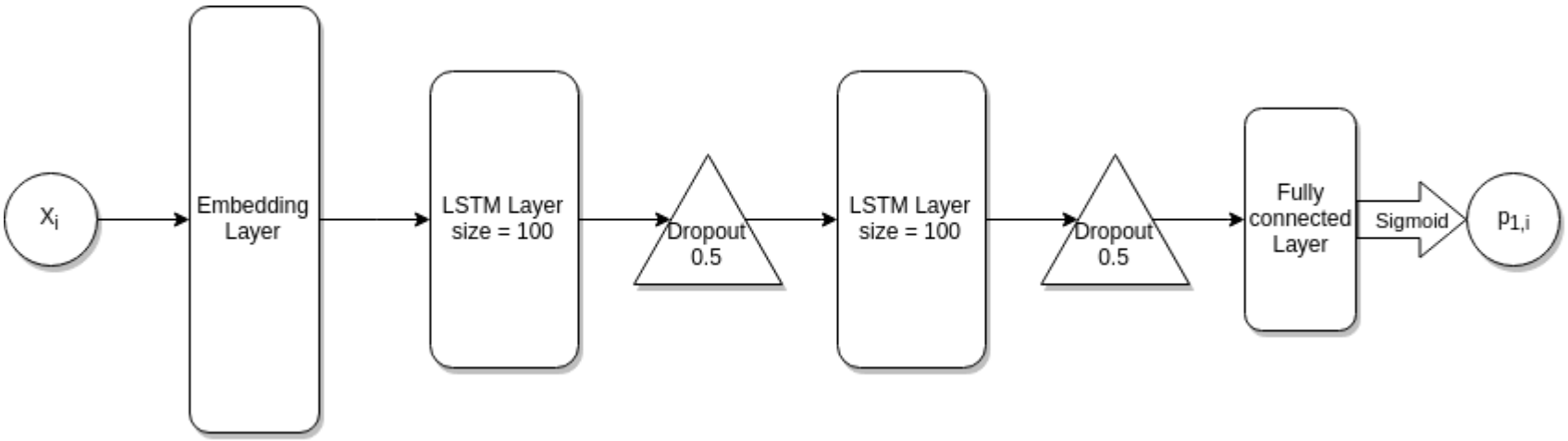
richer model, individual representation for each token, character level representation (no item is OOV), not very practical

1024-dimensional vectors, averaged from three representation vectors

---

- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. NAACL HLT

text classifer



output: 0 to 1 score for text belonging to “at risk” class

output of the text classifier passed as input to author classifier together with

- the number of texts seen
- average score of seen texts
- standard deviation of seen scores
- average score of top 20% highest texts
- difference between average of top 20% and bottom 20% texts

output of the text classifier passed as input to author classifier together with

- the number of texts seen
- average score of seen texts
- standard deviation of seen scores
- average score of top 20% highest texts
- difference between average of top 20% and bottom 20% texts

output of the text classifier passed as input to author classifier together with

- the number of texts seen
- average score of seen texts
- standard deviation of seen scores
- average score of top 20% highest texts
- difference between average of top 20% and bottom 20% texts

output of the text classifier passed as input to author classifier together with

- the number of texts seen
- average score of seen texts
- standard deviation of seen scores
- average score of top 20% highest texts
- difference between average of top 20% and bottom 20% texts

output of the text classifier passed as input to author classifier together with

- the number of texts seen
- average score of seen texts
- standard deviation of seen scores
- average score of top 20% highest texts
- difference between average of top 20% and bottom 20% texts

output of the text classifier passed as input to author classifier together with

- the number of texts seen
- average score of seen texts
- standard deviation of seen scores
- average score of top 20% highest texts
- difference between average of top 20% and bottom 20% texts

author classifier

- logistic regression (linear)
- multi-layer perceptron (non-linear)

author classifier

- logistic regression (linear)
- multi-layer perceptron (non-linear)

author classifier

- logistic regression (linear)
- multi-layer perceptron (non-linear)

author classifier

- logistic regression (linear)
- multi-layer perceptron (non-linear)

hyperparameter (dimensionality, etc) optima different for different representations. compromises were made.

evaluation

- precision and recall calculated over only positive items
- ERDE factors in latency cost factor, penalising late decisions

- implementation using Sci-kit Learn, Keras, NLTK
- some preprocessing of e.g. URLs and numbers; blank texts discarded
- experimentation and hyperparameter setting based on discussions about use case validity (how early do we need to be?)
- more details in paper

- implementation using Sci-kit Learn, Keras, NLTK
- some preprocessing of e.g. URLs and numbers; blank texts discarded
- experimentation and hyperparameter setting based on discussions about use case validity (how early do we need to be?)
- more details in paper

- implementation using Sci-kit Learn, Keras, NLTK
- some preprocessing of e.g. URLs and numbers; blank texts discarded
- experimentation and hyperparameter setting based on discussions about use case validity (how early do we need to be?)
- more details in paper

- implementation using Sci-kit Learn, Keras, NLTK
- some preprocessing of e.g. URLs and numbers; blank texts discarded
- experimentation and hyperparameter setting based on discussions about use case validity (how early do we need to be?)
- more details in paper

- implementation using Sci-kit Learn, Keras, NLTK
- some preprocessing of e.g. URLs and numbers; blank texts discarded
- experimentation and hyperparameter setting based on discussions about use case validity (how early do we need to be?)
- more details in paper

official results

... were broken. disregard them.

instead

		Base	RI	GloVe	ELMo	Top eRisk
linear	Precision	0.35	0.34	0.4	0.41	0.77
	Recall	0.89	0.89	0.9	0.9	1.0
	ERDE 5	0.040	0.042	0.045	0.036	0.06
	ERDE 50	0.027	0.026	0.024	0.023	0.03
nonlinear	Precision	0.49	0.64	0.68	0.65	0.77
	Recall	0.77	0.63	0.67	0.7	1.0
	ERDE 5	0.051	0.065	0.070	0.068	0.06
	ERDE 50	0.035	0.045	0.043	0.038	0.03

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?

lessons

1. continued experiments match the best official submitted results
2. able to include ELMo only after boss fight
3. choice of representation and classifier has some effect on result
4. pretrained vectors gave small benefit
5. learning curve is a factor given ERDE (RI)
6. high recall gives high ERDE (GloVE+MLP; ELMo+LR)
7. more conservative models are likely to perform better in long run but dare not make early decision
8. continuous leaderboard, maybe?