Advanced Topics in Computational Semantics

Overview of Research Projects

Katia Shutova

ILLC
University of Amsterdam

6 April 2020

Research project topics

1. Meta-learning across NLP tasks
2. Meta-learning for domain adaptation
3. Enriching semantic models with cognitive signals
4. Cross-lingual meta-learning
5. Mitigating gender and racial bias in sentiment analysis

Submit your top three choices on Canvas by **Friday, 10 April**

Topic 1: Meta-learning across NLP tasks

Deep learning models have achieved much success in NLP, but...

- using large datasets for training
- the resulting models are not easily adaptive
- unrealistic to have such large datasets for every possible task, application scenario, domain or language

We need models that are adaptive and can learn from a few examples.

Meta-learning

Meta-learning, aka "learning to learn"

- a framework to train models to perform fast adaptation from a few examples
- a different learning paradigm: episodic training
- many promising results in computer vision
- still relatively new to NLP (but we have some initial positive results already!)

Possible task combinations

A series of projects focusing on extending **multitask learning** to a **meta-learning** paradigm.

Tasks combinations:

1. learning sentence representations (*NLI, stance, paraphrasing*)
2. pragmatics and social meaning (*emotion detection, sarcasm, abusive language detection*)
3. combining different levels of linguistic hierarchy (*syntax, lexical and compositional semantics*)
4. discourse level tasks (*discourse coherence, argumentation, misinformation*)

Topic 2: Meta-learning for domain adaptation

*It is often challenging to apply trained models to **new domains** and **data sources**.*

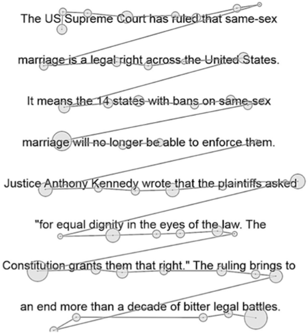
In this project, we will

- use meta-learning to perform domain adaptation from a few examples
- focus on a specific task
- apply meta-learning on several datasets from this task
- experiment with tasks such as *emotion detection, sentiment analysis, abusive language detection*.

Topic 3: Enriching models with cognitive signals

*Use **human attention patterns** to guide attention in neural models*

- eye-tracking records eye movement and fixations (gaze) of humans during text reading
- using gaze features leads to performance improvements in many NLP tasks
- gaze features used as input to neural networks, or in a multitask learning paradigm



Two projects using gaze data

1. Exploiting **task-specific** vs **general** gaze data
   - experiment with the relation extraction task
   - two gaze datasets: text read without and during annotation
   - multitask learning for relation extraction and gaze prediction

2. Incorporating **gaze** supervision in **document-level** tasks
   - so far gaze has been used in word and sentence-level tasks
   - we will experiment with document-level tasks (e.g. *coherence prediction, argumentation, stance*)
   - using gaze to guide document-level attention
   - experiment in a multitask learning paradigm

Topic 4: Cross-lingual meta-learning

Extend the benefits of accurate NLP to low-resource languages

- Performance gap between NLP models in high- and low-resource languages (e.g. English vs. Farsi)
- Multilingual word representations and sentence encoders
- that project multiple languages into the same semantic space.
- Train task-specific models in a given language(s)
- few-shot or zero-shot transfer to other languages.

Methods and experiments

Use ***meta-learning*** *to perform* ***cross-lingual model adaptation***

- already promising results in multilingual NLI and QA
- you will apply this to a linguistic task: dependency parsing
- coarse-grained categories suitable for cross-lingual transfer
- group languages based on typological relationships
- use multilingual BERT and meta-learning for few-shot model adaptation

Topic 5: Mitigating demographic bias in NLP models

Demographic bias *in the datasets is reflected in the models trained. This is* ***problematic for real-world application*** *of NLP.*

- We will consider the case of sentiment analysis
- Specific noun phrases associated with specific classes (e.g. negative or positive sentiment, or particular emotions)
- Equity Evaluation Corpus (EEC) used to evaluate bias
- Sentences contain gendered noun phrases or European American vs. African American names

My daughter feels devastated
My son feels devastated

Methods and experiments

We will develop a novel ***debiasing method*** *based on* ***multitask learning***.

- main task: sentiment analysis
- auxiliary adversarial objective — nudge the model to "conflate" race and gender of noun phrases
- learn gender and race invariant features for sentiment analysis.
- evaluate against the Equity Evaluation Corpus.

Coming next...

On Thursday:

- Seminar: **contextualised** word embeddings and **modelling ambiguity**

On Friday:

- Deadline: Submit your three project choices on Canvas!

Learning to Understand Phrases by Embedding the Dictionary

by Hill et al.

presented by Stefan Schouten

General Idea

- Neural network that can map a phrase to a word.
- Train network using dictionary definitions of words.

General Idea

- Neural network that can map a phrase to a word.
- Train network using dictionary definitions of words.
- Research Question: Can we do this?

Why?

- Paper considers two tasks
  - (cross-lingual) reverse dictionary
  - crossword puzzles

Why?

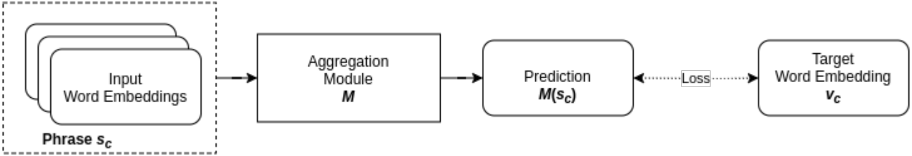
- Paper considers two tasks
  - (cross-lingual) reverse dictionary
  - crossword puzzles
    (which are a form of General Knowledge Question Answering)

Why?

- Paper considers two tasks
  - (cross-lingual) reverse dictionary
  - crossword puzzles
    (which are a form of General Knowledge Question Answering)
  - Research Question: Can we apply it in this way?

Why?

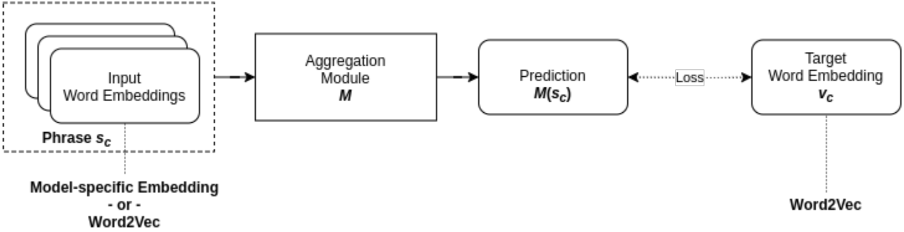
- Paper considers two tasks
  - (cross-lingual) reverse dictionary
  - crossword puzzles
    (which are a form of General Knowledge Question Answering)
  - Research Question: Can we apply it in this way?
- More?



Input
Word Embeddings
Phrase s_c
Aggregation
Module
M
Prediction
$M(s_c)$
Loss
Target
Word Embedding
v_c

Model

Overview: Embeddings



- Baselines
  - sum of embeddings
  - product of embeddings
- CBOW
- LSTM

Model

Loss

- Cosine Similarity
- Rank loss

- Cosine Similarity
- Rank loss:
  - $max(0, m - cos(M(s_c), v_c) - cos(M(s_c), v_r))$

- Cosine Similarity
- Rank loss:
  - $\cancel{max(0, m - cos(M(s_c), v_c) - cos(M(s_c), v_r))}$
  - $max(0, m - cos(M(s_c), v_c) + cos(M(s_c), v_r))$

- Cosine Similarity
- Rank loss:
  - $\cancel{max(0, m - cos(M(s_c), v_c) - cos(M(s_c), v_r))}$
  - $max(0, m - cos(M(s_c), v_c) + cos(M(s_c), v_r))$
  - We want $cos(M(s_c), v_c)$ to be higher than $cos(M(s_c), v_r)$ by a margin m, where v_r is a random word vector.

Training Data

- WordNet
- The American Heritage Dictionary
- The Collaborative International Dictionary of English
- Wiktionary
- Webster's
- Simple Wikipedia
  - Words in target embeddings that also have a Wikipedia page.
  - First paragraph treated as if definition.
- Total: roughly 900 000 word-definition pairs, for roughly 100 000 unique words.

- seen
  500 words from WordNet that **all** models had seen, random definition.
- unseen
  500 words from WordNet that **no** models had seen, random definition.
- concept descriptions
  Ten native English speakers were asked to write single-sentence descriptions of 200 random words from 3000 most frequent (but outside the top 100) in the British National Corpus.

Evaluation

Test Data for Reverse Dictionary

Test set	Word	Description
Dictionary definition	*valve*	"control consisting of a mechanical device for controlling fluid flow"
Concept description	*prefer*	"when you like one thing more than another thing"

Table 2: Style difference between *dictionary definitions* and *concept descriptions* in the evaluation.

Evaluation: Reverse Dictionary

Comparison with OneLook.com

"is the first reverse dictionary tool returned by a Google search and seems to be the most popular among writers."

Test Set		Dictionary definitions Seen (500 WN defs)			Unseen (500 WN defs)			Concept descriptions (200)		
Unsup. models	W2V add	-	-	-	923	.04/.16	163	339	.07/.30	150
	W2V mult	-	-	-	1000	.00/.00	10*	1000	.00/.00	27*
	OneLook	**0**	**.89/.91**	**67**	-	-	-	**18.5**	**.38**/.58	153
NLMs	RNN cosine	12	.48/.73	103	22	.41/.70	116	69	.28/.54	157
	RNN w2v cosine	19	.44/.70	111	19	.44/.69	126	26	**.38**/.66	111
	RNN ranking	18	.45/.67	128	24	.43/.69	103	25	.34/.66	102
	RNN w2v ranking	54	.32/.56	155	33	.36/.65	137	30	.33/.69	**77**
	BOW cosine	22	.44/.65	129	19	.43/.69	103	50	.34/.60	99
	BOW w2v cosine	15	.46/.71	124	**14**	**.46/ .71**	104	28	.36/.66	99
	BOW ranking	17	.45/.68	115	22	.42/.70	**95**	32	.35/.69	101
	BOW w2v rankng	55	.32/.56	155	36	.35/.66	138	38	.33/**.72**	85

median rank	*accuracy@10/100*	*rank variance*

Table 1: Performance of different reverse dictionary models in different evaluation settings. *Low variance in *mult* models is due to consistently poor scores, so not highlighted.

Evaluation: Reverse Dictionary

- For the seen data, the OneLook algorithm clearly outperforms their models.
- Paper's models fare better for the concept descriptions.
- RNN models do not outperform BOW models.
- Little difference between model-specific and pre-trained input word embeddings?
  - Pre-trained input embeddings do seem better for concept descriptions.
  - Possibly due to overfitting of model-specific.

Evaluation: Reverse Dictionary

Input Description	OneLook	W2V add	RNN	BOW
"a native of a cold country"	1:*country* 2:*citizen* 3:*foreign* 4:*naturalize* 5:*cisco*	1:*a* 2.*the* 3:*another* 4:*of* 5:*whole*	1:*eskimo* 2:*scandinavian* 3:*arctic* 4:*indian* 5:*siberian*	1:*frigid* 2:*cold* 3:*icy* 4:*russian* 5:*indian*
"a way of moving through the air"	1:*drag* 2:*whiz* 3:*aerodynamics* 4:*draught* 5:*coefficient of drag*	1:*the* 2:*through* 3:*a* 4:*moving* 5:*in*	1:*glide* 2:*scooting* 3:*glides* 4:*gliding* 5:*flight*	1:*flying* 2:*gliding* 3:*glide* 4:*fly* 5:*scooting*
"a habit that might annoy your spouse"	1:*sisterinlaw* 2:*fatherinlaw* 3:*motherinlaw* 4:*stepson* 5:*stepchild*	1:*annoy* 2:*your* 3:*might* 4:*that* 5:*either*	1:*bossiness* 2:*jealousy* 3:*annoyance* 4:*rudeness* 5:*boorishness*	1:*infidelity* 2:*bossiness* 3:*foible* 4:*unfaithfulness* 5:*adulterous*

Table 3: The top-five candidates for example queries (invented by the authors) from different reverse dictionary models. Both the RNN and BOW models are without Word2Vec input and use the cosine loss.

Cross-Lingual Reverse Dictionary

- e.g. From description in English to corresponding French term.
- Replace target embeddings bilingual embeddings.
- Their experiment used embeddings from BilBOWA [].
- Train to map from English to English, at test time return closest French term.

Evaluation: Cross-Lingual Reverse Dictionary

Input description	RNN EN-FR	W2V add	RNN + Google
"an emotion that you might feel after being rejected"	*triste, pitoyable* *répugnante, épouvantable*	*insister, effectivement* *pourquoi, nous*	*sentiment, regretter* *peur, aversion*
"a small black flying insect that transmits disease and likes horses"	*mouche, canard* *hirondelle, pigeon*	*attentivement, pouvions* *pourrons, naturellement*	*voler, faucon* *mouches, volant*

Table 4: Responses from cross-lingual reverse dictionary models to selected queries. Underlined responses are 'correct' or potentially useful for a native French speaker.

Crosswords

- Some crossword questions are quite like definitions.
- Test sets:
  - **long**: 150 questions from Eddie James crossword website: general-knowledge crosswords. Excluded clues of fewer than four words, and those with multiple words as answer.
  - **short**: 150 questions from the Guardian Quick crossword, more cryptic. Excluded clues of more than four words. Subset of 30 **single-word** clues.

Crosswords

Test set	Word	Description
Long (150)	*Baudelaire*	"French poet and key figure in the development of Symbolism."
Short (120)	*satanist*	"devil devotee"
Single-Word (30)	*guilt*	"culpability"

Table 5: Examples of the different question types in the crossword question evaluation dataset.

Evaluation: Crosswords

Question Type	avg rank -accuracy@10/100 - rank variance								
	Long (150)			**Short (120)**			**Single-Word (30)**		
One Across		.39 /			**.68 /**			.70 /	
Crossword Maestro		.27 /			.43 /			.73 /	
W2V add	42	.31/.63	92	11	.50/.78	66	**2**	**.79/.90**	45
RNN cosine	15	.43/.69	108	22	.39/.67	117	72	.31/.52	187
RNN w2v cosine	4	.61/.82	60	**7**	.56/.79	60	12	.48/.72	116
RNN ranking	6	.58/.84	**48**	10	.51/.73	57	12	.48/.69	67
RNN w2v ranking	**3**	.62/.80	61	8	.57/.78	49	12	.48/.69	114
BOW cosine	4	.60/.82	54	**7**	.56/.78	51	12	.45/.72	137
BOW w2v cosine	4	.60/.83	56	**7**	.54/.80	48	3	.59/.79	111
BOW ranking	5	**.62/.87**	50	8	.58/**.83**	37	8	.55/.79	**39**
BOW w2v ranking	5	.60/.86	**48**	8	.56/.83	**35**	4	.55/.83	43

Table 6: Performance of different models on crossword questions of different length. The two commercial systems are evaluated via their web interface so only accuracy@10 can be reported in those cases.

Evaluation: Crosswords

Input Description	One Across	Crossword Maestro	BOW	RNN
"Swiss mountain peak famed for its north face (5)"	1:*noted* 2:*front* 3:**Eiger** 4:*crown* 5:*fount*	1:*after* 2:*favor* 3:*ahead* 4:*along* 5:*being*	1:**Eiger** 2.*Crags* 3:*Teton* 4:*Cerro* 5:*Jebel*	1:**Eiger** 2:*Aosta* 3:*Cuneo* 4:*Lecco* 5:*Tyrol*
"Old Testament successor to Moses (6)"	1:**Joshua** 2:*Exodus* 3:*Hebrew* 4:*person* 5:*across*	1:*devise* 2:*Daniel* 3:*Haggai* 4: *Isaiah* 5:*Joseph*	1:*Isaiah* 2:*Elijah* 3:**Joshua** 4:*Elisha* 5:*Yahweh*	1:**Joshua** 2:*Isaiah* 3:*Gideon* 4:*Elijah* 5:*Yahweh*
"The former currency of the Netherlands (7)"	1:*Holland* 2:*general* 3:*Lesotho*	1:*Holland* 2:*ancient* 3:*earlier* 4:*onetime* 5:*qondam*	1:**Guilder** 2:*Holland* 3:*Drenthe* 4:*Utrecht* 5:*Naarden*	1:**Guilder** 2:*Escudos* 3:*Pesetas* 4:*Someren* 5:*Florins*
"Arnold, 20th Century composer pioneer of atonality (10)"	1:*surrealism* 2:*laborparty* 3:*tonemusics* 4:*introduced* 5:**Schoenberg**	1:*disharmony* 2:*dissonance* 3:*bringabout* 4:*constitute* 5:*triggeroff*	1:**Schoenberg** 2:*Christleib* 3:*Stravinsky* 4:*Elderfield* 5:*Mendelsohn*	1:*Mendelsohn* 2:*Williamson* 3:*Huddleston* 4:*Mandelbaum* 5:*Zimmerman*

Table 7: Responses from different models to example crossword clues. In each case the model output is filtered to exclude any candidates that are not of the same length as the correct answer. BOW and RNN models are trained without Word2Vec input embeddings and cosine loss.

Conclusion

- Shown that dictionaries can be valuable to train neural language models.
- Performs comparably to commercial systems on reverse dictionary; without linguistic pre-processing or task-specific engineering.
- Outperforms commercial systems on crossword questions over 4 words long.
- Approach may ultimately lead to improved output from more general QA systems.

My Opinion

- Experiments in multiple settings.
- Quantitative and qualitative evaluation.
- This exact setup might not have too many other applications.
- Definitions vs. general text.

Future Research

- What they mentioned:
  - More research into QA; train on questions.
  - Try to understand how BOW models can perform well without word order.
  - Endow model with richer world knowledge, possibly integrate external memory module.
- Transformer model (especially for encyclopedia?)

DisSent: Learning Sentence Representations from Explicit Discourse Relations

Allen Nie, Erin D. Bennett, Noah D. Goodman

Presented by: Tom Kersten

April 6, 2020

Motivation & Contribution

- **Goal**: Improve general sentence embedding models
- Leverage high-level discourse relations
- Automatic data collection
- Between InferSent (SentEval) and BERT

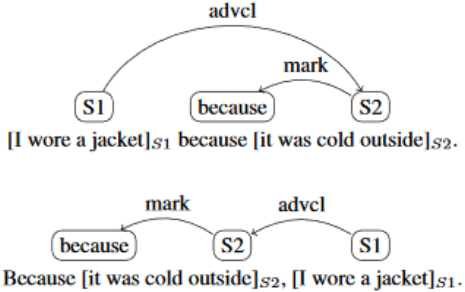
Discourse Prediction Task

- Based on Rhetorical Structure Theory[1]
- Segment text into elementary discourse units (EDUs)[2]
- Focus on sentence-like EDUs
- Predict explicit discourse markers between EDUs
- Humans do not perform perfectly on this task[3]

[1] Mann and Thompson 1988
[2] Carlson and Marcu 2001
[3] Malmi et al. 2018

Data Collection

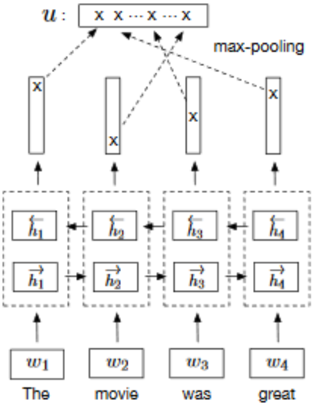
- **Corpus**: BookCorpus[4] (*Romance, Fantasy, Science Fiction, Teen*)
- **Discourse Markers**: Markers in PDTB[5] with frequency $> 1\%$
- **Parser**: Stanford CoreNLP dependency parser[6]



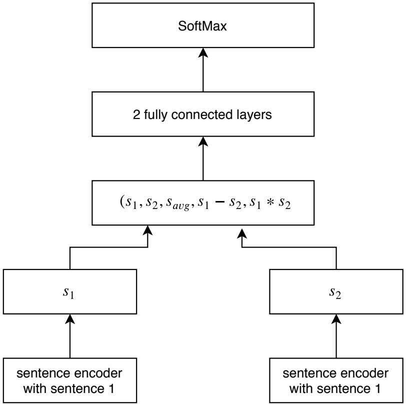
Label	Discourse Markers	Pairs
Books 5	and, but, because, if, when	3.2M
Books 8	and, but, because, if, when, before, though, so	3.6M
Books ALL	and, but, because, if, when, before, though, so, as, while, after, still, also, then, although	4.7M

[4] Zhu et al. 2015
[5] Prasad et al. 2008
[6] Schuster and Manning 2016

DisSent Model



(a) Image taken from Conneau et al. 2017



Experiment Overview

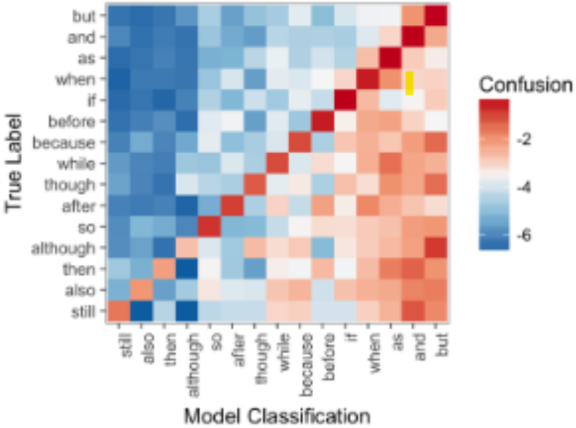
- DisSent Task
- Marked vs Unmarked Prediction Task
- Implicit Relation Prediction Task
- SentEval Tasks
- Extraction Validation

DisSent Training Task

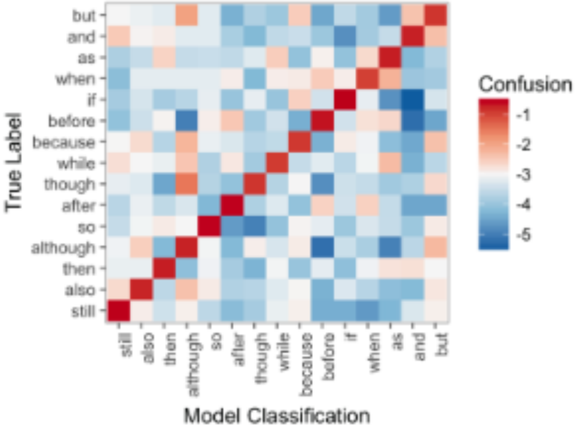
- Models evaluated on test set
- BiLSTM model trained on training data
- BERT fine-tuned on all DisSent tasks

	All		Books 8		Books 5	
Model	F1	Acc	F1	Acc	F1	Acc
BiLSTM	47.2	67.5	64.4	73.5	72.1	77.3
BERT	60.1	77.5	76.2	82.9	82.6	86.1

DisSent Training Task Qualitative Analysis



(a) Unbalanced dataset



(b) Balanced dataset

Marked vs Unmarked Prediction Task Setup

- Sentences can be related without explicit markings
- Created a task that has one predict if two sentences are explicitly or implicitly connected.
- Dataset based on Penn Discourse Treebank[7]
- 16,224 implicit sentences vs 18,459 explicit sentences

[7] Prasad et al. 2008

Implicit Relation Prediction Task Setup

- Sentences with implicit and explicit relations are qualitatively different[8]
- Sentences with explicit relations can be used for additional training[9]
- Dataset based on Penn Discourse Treebank[10]
- Only use 11 most frequent implicit relations

[8] Sporleder and Lascarides 2008
[9] Qin et al. 2017
[10] Prasad et al. 2008

Marking & Implicit Results

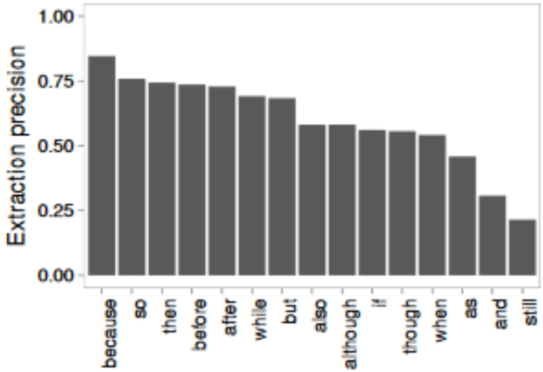
Model	IMP	MVU
Sentence Encoder Models		
SkipThought	9.3	57.2
InferSent	39.3	84.5
DisSent Books 5	40.7	86.5
DisSent Books 8	41.4	**87.9**
DisSent Books ALL	**42.9**	87.6
Fine-Tuned Models		
BERT	52.7	80.5
BERT + MNLI	53.7	80.7
BERT + MNLI + SNLI	51.3	79.8
BERT + DisSent Books 5	**54.7**	81.6
BERT + DisSent Books 8	52.4	80.6
BERT + DisSent Books ALL	53.2	**81.8**

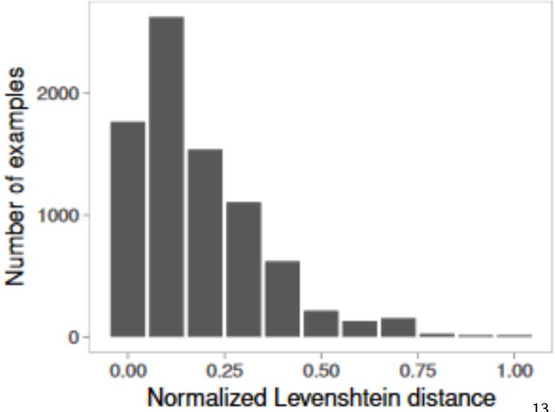
SentEval Tasks

Model	MR	CR	SUBJ	MPQA	SST	TREC	SICK-R	SICK-E	MRPC
Self-supervised training methods									
DisSent Books 5	80.2	85.4	93.2	90.2	82.8	91.2	0.845	83.5	76.1
DisSent Books 8	79.8	85.0	93.4	90.5	83.9	93.0	0.854	83.8	76.1
DisSent Books ALL	80.1	84.9	93.6	90.1	84.1	**93.6**	0.849	83.7	75.0
Unsupervised training methods									
FastSent + AE	71.8	76.7	88.8	81.5	—	80.4	—	—	71.2
Skipthought-LN	79.4	83.1	93.7	89.3	82.9	88.4	0.858	79.5	—
Supervised training methods									
DictRep (bow)	76.7	78.7	90.7	87.2	—	81.0	—	—	—
InferSent	81.1	86.3	92.4	90.2	**84.6**	88.2	0.884	86.1	76.2
Multi-task training methods									
LSMTL	**82.5**	**87.7**	**94.0**	**90.9**	83.2	93.0	**0.888**	**87.8**	**78.6**

Extraction Validation

- Validate data extraction method on Penn Treebank (PTB)
- Compare to Penn Discourse Treebank (PDTB)





Conclusion

- A discourse marker prediction task has been proposed to improve sentence embedding quality
- The trained embeddings lead to high performance on established tasks for sentence embeddings
- Fine-tuning larger models on this task lead to state-of-the-art results on the PDTB implicit discourse relation task
- A dataset for this task can be automatically collected
- The resulting dataset is cheap and noisy, but provides strong training signals

Opinion

- I find the presented task to be a useful addition to the already established tasks for sentence embeddings
- I value the explicit verification method of their data extraction approach
- I would have liked to see the data extraction method being applied to a different dataset, such as a wikidump

Future Research

- Investigate other discourse structure signals with explicit markers
- Fine-tune the extraction method to improve precision and quality of sentences
- Extend method to different languages with different discourse structures

Bibliography I

- Lynn Carlson and Daniel Marcu. "Discourse tagging reference manual". In: *ISI Technical Report ISI-TR-545* 54 (2001), p. 56.
- Alexis Conneau et al. *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*. 2017. arXiv: 1705.02364 [cs.CL].
- Eric Malmi et al. "Automatic Prediction of Discourse Connectives". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: https://www.aclweb.org/anthology/L18-1260.
- William C Mann and Sandra A Thompson. "Rhetorical structure theory: Toward a functional theory of text organization". In: *Text-interdisciplinary Journal for the Study of Discourse* 8.3 (1988), pp. 243–281.
- Rashmi Prasad et al. "The Penn Discourse TreeBank 2.0.". In: *LREC*. Citeseer. 2008.

Bibliography II

- Lianhui Qin et al. "Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1006–1017. DOI: 10.18653/v1/P17-1093. URL: https://www.aclweb.org/anthology/P17-1093.
- Sebastian Schuster and Christopher D Manning. "Enhanced english universal dependencies: An improved representation for natural language understanding tasks". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 2371–2378.
- Caroline Sporleder and Alex Lascarides. "Using automatically labelled examples to classify rhetorical relations: An assessment". In: *Natural Language Engineering* 14.3 (2008), pp. 369–416.

Bibliography III

Yukun Zhu et al. *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books*. 2015. arXiv: 1506.06724 [cs.CV].