# RandSent

MICKUS, Timothee

General layout of this presentation of Wieting et Kiela (2019) :

- A word on sentence embeddings
- No training required ?
- Evaluation results

# Sentence embeddings

# Sentence Embeddings

## How to train a sentence representation ?

- Landmark paper for the ‘opaque’ approach : Skipthought by Kiros et al. (2015)
  - train a GRU seq2seq model to generate the surrounding sentences ; the output of the encoder is the representation for that sentence.
- InferSent (Conneau et al., 2017)
  - consider this problem as twofold : 1/ ‘what architecture ?’, 2/ ‘how to train it ?’ and settle for a BiLSTM trained on natural language inference.
- Quickthought (Logeswaran et Lee, 2018)
  - reformulate Kiros et al. (2015) as a classification problem (do encoded sentences appear in the same context ?) rather than a generation problem (knowing sentence, generate context).
- Universal Sentence Encoder (Cer et al., 2018)
  - train a Transformer in a multi-task framework, including NLI, context generation and conversation-based generation (also test a deep averaging network).

# Sentence Embeddings

## How to train a sentence representation ?

Another approach focuses on mimicking compositional semantics so as to make composition 'transparent' :

- Baroni et Zamparelli (2010) suggest that "words are vectors, adjectives are matrices ; ie. of different types.
- Grefenstette et al. (2013) propose to use tensors and vectors to model functions and arguments
- Paperno, Pham et Baroni (2014) treat functions of arity $n$ as tensors of rank $n$, and learn different representations for each words based on their usage
- Hill et al. (2016) et Hill, Cho et Korhonen (2016) use dictionaries to propose a rational supervised objective for composition ; thus taking a middle ground between the 'opaque' and 'transparent' approaches

# Sentence Embeddings

## Composition ?

- All these architectures entail that you need a “composition function”.
- however compositional distributional semantics is still an open problem (Lenci, 2018, ao.) ; simple summations or element-wise products (Mitchell et Lapata, 2008) don’t really cut it
- As a side note, these models also questions linguistic theories on composition, as they try to combine distributional semantics (Wittgenstein, 1921 ; Quine, 1960) with Frege’s compositionality principle (Frege, 1892).

# Sentence Embeddings

## Analyzing sentence encoders

- Circumventing all that, recent work has focused on putting sentence representations to the test
  - most notable is **SentEval** (Conneau et Kiela, 2018) which lists an array of tasks with which sentence representations should help.
  - another popular benchmark is **GLUE** (Wang et al., 2018), the "General Language Understanding Evaluation benchmark"
  - Adi et al. (2016) propose to train classifiers on low-level sentence properties (length, word content, word order...)
  - Conneau et al. (2018) suggest to probe sentence representations for linguistic properties
  - Linzen, Dupoux et Goldberg (2016) study whether sentence encoder are able to work out long-term syntactic dependencies like agreement the way humans do.
- Today's paper is related to this last trend : Wieting et Kiela (2019) study whether sentence encoders do better than **randomly initialized architectures**.

## How random is a sentence encoder ?

# Random encoders

The main insight is drawn from Cover (1965) :

> *A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated*

Therefore, to evaluate sentence encoders, one needs to tease apart what is to be attributed to the nonlinear high-dimension projection from what is due to the training regimen

# Random encoders

## Why so random ?

Teasing these two factors apart is crucial for multiple reasons :

- Sentence encoders require extensive resources for training
- Raw word-embeddings with simple pooling mechanisms already perform quite well (Shen et al., 2018)
- Such studies shed light on whether some architectures are sounder models than others

# Random encoders

## What do you mean, random ?

The proposed methodology is the following :

- take pre-trained word embeddings
- initialize a composition function, **don't** (fully) **train it**
- train a linear classifier on top of the untrained composed representation for each senteval task
- compare results

# Random encoders

## Random models

Wieting et Kiela (2019) suggest two random models, with values initialized in $[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]$ where $d$ is the size of the input embeddings :

1. a random linear transformation (BOREP : "bag of random embedding projections") : $h_i = We_i$, with an optional reLU non-linearity $(\max(0, h))$
2. a BiLSTM, as used in InferSent (Conneau et al., 2017)

as well as three 'pooling' mechanisms

1. summation : $\vec{\text{sentence}} = \sum h$
2. max-pooling : $\vec{\text{sentence}} = \max(h)$
3. mean-pooling : $\vec{\text{sentence}} = |h|^{-1} \sum h$

# Random encoders

## Random models

- Wieting et Kiela (2019) also study a bidirectional Echo State Network (Jaeger, 2001), which assigns a representation $\hat{y}_i$ to each embedding $e_i$ of a sequence based on a gating mechanism :

$$\tilde{h}_i = \text{pool}(W^i e + W^h h_{i-1} + b^i)$$

$$h_i = (1 - \alpha) h_{i-1} + \alpha \tilde{h}_i$$

$$\hat{y}_i = W^o (e_i \oplus h_i) + b^o$$

- Contrarily to the two previous models, the output linear transformation parameters $W^o$ and $b^o$ are learned. A sentence representation is derived using max-pooling over all predicted outputs $\hat{y}_i$.
- An additional property, called ‘echo state property’, required of this model is that the intermediary representations $h_i$ must be uniquely determined by the input history of the ESN. This property is guaranteed by specific initialization procedures for $W^h$ and $W^i$

# Evaluation Results

# Results

## Recap

- Wieting et Kiela (2019) are teasing apart effects of high-dimensionality projection and training procedures
- They compare results on SentEval (Conneau et Kiela, 2018), by training a classifier on top of sentence representations
- They compare randomly or partially randomly initialized models to existing sentence encoders, namely Skipthought (Kiros et al., 2015) and InferSent (Conneau et al., 2017)

# Results

## SentEval

SentEval (Conneau et Kiela, 2018) is a benchmark composed of multiple sentence-level tasks :

- Sentiment analysis : **MR** (movie reviews, Pang et Lee (2005)), **CR** (customer reviews, Hu et Liu (2004)), **MPQA** (opinion polarity, Wiebe, Wilson et Cardie (2005)), and **SST** (movie review, Socher et al. (2013))
- Semantic properties : **TREC** (question type, Voorhees et Tice (2000)), **STSB** (relatedness, Cer et al. (2017)), **SICK-R** (relatedness, Marelli et al. (2014)), **SICK-E** (entailment, Marelli et al. (2014)), **SNLI** (entailment, Bowman et al. (2015)), **SUBJ** (subjectivity/objectivity classification, Pang et Lee (2004)) and **MRPC** (paraphrases, Dolan, Quirk et Brockett (2004))

# Results

## Comparing on 4096 dimensions

| Model | Dim | MR | CR | MPQA | SUBJ | SST2 | TREC | SICK-R | SICK-E | MRPC | STSB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BOE | 300 | 77.3(.2) | 78.6(.3) | 87.6(.1) | 91.3(.1) | 80.0(.5) | 81.5(.8) | 80.2(.1) | 78.7(.1) | 72.9(.3) | 70.5(.1) |
| BOREP | 4096 | 77.4(.4) | 79.5(.2) | 88.3(.2) | 91.9(.2) | 81.8(.4) | **88.8(.3)** | 85.5(.1) | 82.7(.7) | 73.9(.4) | 68.5(.6) |
| RandLSTM | 4096 | 77.2(.3) | 78.7(.5) | 87.9(.1) | 91.9(.2) | 81.5(.3) | 86.5(1.1) | 85.5(.1) | 81.8(.5) | **74.1(.5)** | 72.4(.5) |
| ESN | 4096 | **78.1(.3)** | **80.0(.6)** | **88.5(.2)** | **92.6(.1)** | **83.0(.5)** | 87.9(1.0) | **86.1(.1)** | **83.1(.4)** | 73.4(.4) | **74.4(.3)** |
| InferSent-1 = paper, G | 4096 | 81.1 | 86.3 | 90.2 | 92.4 | 84.6 | 88.2 | 88.3 | 86.3 | 76.2 | 75.6 |
| InferSent-2 = fixed pad, F | 4096 | 79.7 | 84.2 | 89.4 | 92.7 | 84.3 | 90.8 | 88.8 | 86.3 | 76.0 | 78.4 |
| InferSent-3 = fixed pad, G | 4096 | 79.7 | 83.4 | 88.9 | 92.6 | 83.5 | 90.8 | 88.5 | 84.1 | 76.4 | 77.3 |
| Δ InferSent-3, BestRand | - | *1.6* | *3.4* | *0.4* | *0.0* | *0.5* | *2.0* | *2.4* | *1.0* | *2.3* | *2.9* |
| ST-LN | 4800 | 79.4 | 83.1 | 89.3 | 93.7 | 82.9 | 88.4 | 85.8 | 79.5 | 73.2 | 68.9 |
| Δ ST-LN, BestRand | - | *1.3* | *3.1* | *0.8* | *1.1* | *-0.1* | *0.5* | *-0.3* | *-3.6* | *-0.9* | *-5.5* |

Table 1: Performance (accuracy for all tasks except SICK-R and STSB, for which we report Pearson's $r$) on all ten downstream tasks where all models have 4096 dimensions with the exception of BOE (300) and ST-LN (4800). Standard deviations are show in parentheses. InferSent-1 is the paper version with GloVe (G) embeddings, InferSent-2 has fixed padding and uses FastText (F) embeddings, and InferSent-3 has fixed padding and uses GloVe embeddings. We also show the difference between the best random architecture (BestRand) and InferSent-3 and ST-LN, respectively. The average performance difference between the best random architecture and InferSent-3 and ST-LN is 1.7 and -0.4 respectively.

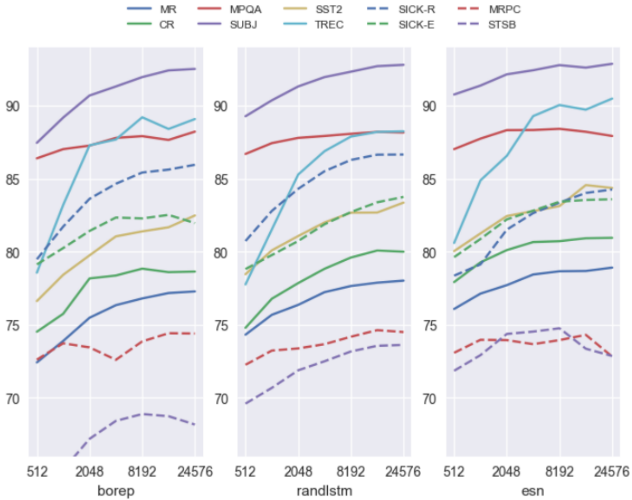
# Results

## Comparing on 4096 dimensions

- We already see that random models constitute a strong baseline :
  - In other words, InferSent does not improve that much over random models
  - On average, SkipThought performs worse than random models.
- Lower results from SkipThought might be due to the fact it uses older embeddings,
- Higher results from ESN might be due to the wider number of hyperparameters available
- what happens for even higher dimensionalities ?

# Results

## Comparing on even more dimensions



# Results

## Comparing on even more dimensions

| Model | MR | CR | MPQA | SUBJ | SST2 | TREC | SICK-R | SICK-E | MRPC | STSB |
|---|---|---|---|---|---|---|---|---|---|---|
| BOE | 77.3(.2) | 78.6(.3) | 87.6(.1) | 91.3(.1) | 80.0(.5) | 81.5(.8) | 80.2(.1) | 78.7(.1) | 72.9(.3) | 70.5(.1) |
| BOREP | 78.6(.2) | 79.9(.4) | 88.8(.1) | 93.0(.1) | 82.5(.8) | 89.5(1.3) | 85.9(.0) | 84.3(.3) | 73.7(.9) | 68.3(.5) |
| RandLSTM | 78.2(.2) | 79.9(.4) | 88.2(.2) | 92.8(.2) | 83.2(.4) | 88.4(.7) | 86.6(.1) | 83.0(.9) | 74.7(.4) | **73.6(.4)** |
| ESN | **79.1(.2)** | **80.2(.3)** | **88.9(.1)** | **93.4(.2)** | **84.6(.5)** | **92.2(.8)** | **87.2(.1)** | **85.1(.2)** | **75.3(.6)** | 73.1(.2) |
| InferSent-3 4096×6 | **79.7** | **83.9** | **89.1** | **92.8** | **82.4** | **90.6** | 79.5 | **85.9** | **75.1** | **75.0** |
| ST-LN 4096×6 | 75.2 | 80.8 | 86.8 | 92.7 | 80.6 | 88.4 | **82.9** | 81.3 | 71.5 | 67.0 |

Table 2: Performance (accuracy for all tasks except SICK-R and STSB, for which we report Pearson's $r$) on all ten downstream tasks. Standard deviations are show in parentheses. All models have 4096×6 dimensions. ST-LN and InferSent-3 were projected to this dimension with a random projection.

- Performance of random models increases with dimensionality
- Random projections of sentence encoders is detrimental to their performance

# Recap

# Recap

## Random systems are a very strong baseline

- Dimensionality matters, especially to downstream classifications
- Random projections of embeddings and random initialization are strong and cheap baselines

There's more in the paper : the authors also tested their random models on the probing tasks from Conneau et al. (2018)

# References I


Adi, Yossi et al. (2016). « Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks ». In : *CoRR* abs/1608.04207. arXiv : 1608.04207. url : http://arxiv.org/abs/1608.04207.

Baroni, Marco et Roberto Zamparelli (2010). « Nouns Are Vectors, Adjectives Are Matrices : Representing Adjective-noun Constructions in Semantic Space ». In : *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Cambridge, Massachusetts : Association for Computational Linguistics, p. 1183–1193. url : http://dl.acm.org/citation.cfm?id=1870658.1870773.

Bowman, Samuel R. et al. (2015). « A large annotated corpus for learning natural language inference ». In : *CoRR* abs/1508.05326. arXiv : 1508.05326. url : http://arxiv.org/abs/1508.05326.

Cer, Daniel et al. (2018). « Universal Sentence Encoder ». In : *CoRR* abs/1803.11175. arXiv : 1803.11175. url : http://arxiv.org/abs/1803.11175.

Cer, Daniel M. et al. (2017). « SemEval-2017 Task 1 : Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation ». In : *CoRR* abs/1708.00055. arXiv : 1708.00055. url : http://arxiv.org/abs/1708.00055.

# References II


Conneau, Alexis et Douwe Kiela (2018). « SentEval : An Evaluation Toolkit for Universal Sentence Representations ». In : *CoRR* abs/1803.05449. arXiv : 1803.05449. url : http://arxiv.org/abs/1803.05449.

Conneau, Alexis et al. (2017). « Supervised Learning of Universal Sentence Representations from Natural Language Inference Data ». In : *CoRR* abs/1705.02364. arXiv : 1705.02364. url : http://arxiv.org/abs/1705.02364.

Conneau, Alexis et al. (2018). « What you can cram into a single vector : Probing sentence embeddings for linguistic properties ». In : *CoRR* abs/1805.01070. arXiv : 1805.01070. url : http://arxiv.org/abs/1805.01070.

Cover, Thomas M. (1965). « Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition ». In : *IEEE Trans. Electronic Computers* 14.3, p. 326–334. doi : 10.1109/PGEC.1965.264137. url : https://doi.org/10.1109/PGEC.1965.264137.

Dolan, Bill, Chris Quirk et Chris Brockett (2004). « Unsupervised Construction of Large Paraphrase Corpora : Exploiting Massively Parallel News Sources ». In : *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*. url : http://aclweb.org/anthology/C04-1051.

Frege, Gottlob (1892). « Über Sinn und Bedeutung ». In : *Zeitschrift für Philosophie und philosophische Kritik* 100, p. 25–50.

# References III


Grefenstette, Edward et al. (2013). « Multi-Step Regression Learning for Compositional Distributional Semantics ». In : *IWCS*.

Hill, Felix, Kyunghyun Cho et Anna Korhonen (2016). « Learning Distributed Representations of Sentences from Unlabelled Data ». In : *CoRR* abs/1602.03483. arXiv : 1602.03483. url : http://arxiv.org/abs/1602.03483.

Hill, Felix et al. (2016). « Learning to Understand Phrases by Embedding the Dictionary ». In : *Transactions of the Association for Computational Linguistics* 4, p. 17–30. issn : 2307-387X. url : https://transacl.org/ojs/index.php/tacl/article/view/711.

Hu, Minqing et Bing Liu (2004). « Mining and summarizing customer reviews ». In : *KDD*.

Jaeger, Herbert (2001). « The "Echo State" Approach to Analysing and Training Recurrent Neural Networks ». In : *GMD-Report 148, German National Research Institute for Computer Science*.

Kiros, Ryan et al. (2015). « Skip-Thought Vectors ». In : *CoRR* abs/1506.06726. arXiv : 1506.06726. url : http://arxiv.org/abs/1506.06726.

Lenci, Alessandro (2018). « Distributional models of word meaning ». In : *Annual review of Linguistics* 4, p. 151–171.

# References IV

Linzen, Tal, Emmanuel Dupoux et Yoav Goldberg (2016). « Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies ». In : *CoRR* abs/1611.01368. arXiv : 1611.01368. url : http://arxiv.org/abs/1611.01368.

Logeswaran, Lajanugen et Honglak Lee (2018). « An efficient framework for learning sentence representations ». In : *International Conference on Learning Representations*. url : https://openreview.net/forum?id=rJvJXZb0W.

Marelli, Marco et al. (2014). « A SICK cure for the evaluation of compositional distributional semantic models ». In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland : European Language Resources Association (ELRA). url : http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

Mitchell, Jeff et Mirella Lapata (2008). « Vector-based models of semantic composition ». In : *In Proceedings of ACL-08 : HLT*, p. 236–244.

Pang, Bo et Lillian Lee (2004). « A sentimental education : Sentiment analysis using subjectivity ». In : *Proceedings of ACL*, p. 271–278.

– (2005). « Seeing stars : Exploiting class relationships for sentiment categorization with respect to rating scales ». In : p. 115–124.

# References V


Paperno, Denis, Nghia The Pham et Marco Baroni (2014). « A practical and linguistically-motivated approach to compositional distributional semantics ». In : *ACL*.

Quine, William Van Ormann (1960). *Word And Object*. MIT Press.

Shen, Dinghan et al. (2018). « Baseline Needs More Love : On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms ». In : *CoRR* abs/1805.09843. arXiv : 1805.09843. url : http://arxiv.org/abs/1805.09843.

Socher, R et al. (2013). « Recursive deep models for semantic compositionality over a sentiment treebank ». In : *EMNLP* 1631, p. 1631–1642.

Voorhees, Ellen M. et Dawn M. Tice (2000). « Building a Question Answering Test Collection ». In : *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '00. Athens, Greece : ACM, p. 200–207. isbn : 1-58113-226-3. doi : 10.1145/345508.345577. url : http://doi.acm.org/10.1145/345508.345577.

Wang, Alex et al. (2018). « GLUE : A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding ». In : *CoRR* abs/1804.07461. arXiv : 1804.07461. url : http://arxiv.org/abs/1804.07461.

# References VI


Wiebe, Janyce, Theresa Wilson et Claire Cardie (2005). « Annotating Expressions of Opinions and Emotions in Language ». In : *Language Resources and Evaluation* 1.2, p. 0. url : http://www.cs.pitt.edu/\~{}wiebe/pubs/papers/lre05withappendix.pdf.

Wieting, John et Douwe Kiela (2019). « No Training Required : Exploring Random Encoders for Sentence Classification ». In : *International Conference on Learning Representations*. url : https://openreview.net/forum?id=BkgPajAcY7.

Wittgenstein, Ludwig (1921). *Tractatus Logico-Philosophicus*. Sous la dir. de Wilhelm Ostwald. Annalen der Naturphilosophie, 14.