On the Robustness of Text Vectorizers

Rémi Catellier (Université Côte d'Azur), Samuel Vaiter (CNRS, Université Côte d'Azur), Damien Garreau (Université Côte d'Azur, Inria)

1: Summary

This work = robustness of text vectorization w.r.t. word replacements
Motivation: adversarial examples, certifiability,...
Previous work: does not take into account the *text-to-vector* step (vectorization)

Main result: classical vectorizers are robust to small perturbations of input document

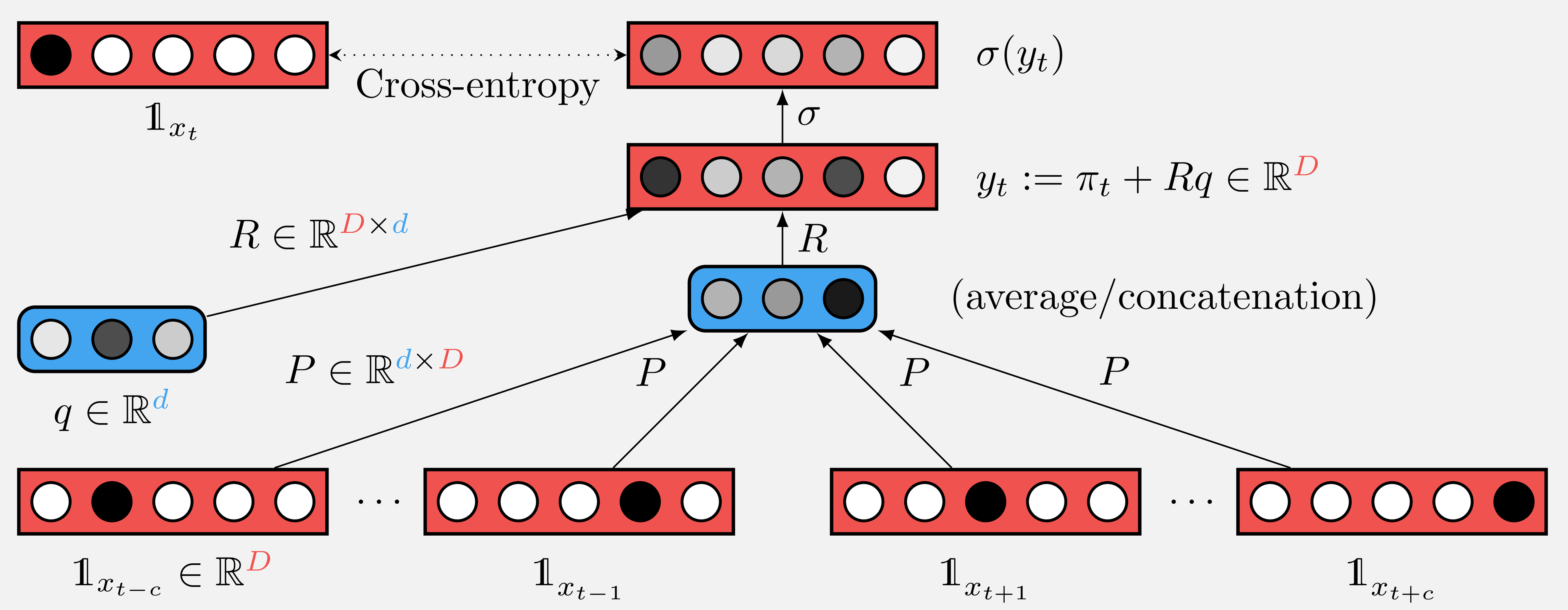
2: Definitions

- **Tokens:** (sub-)words, characters... Belong to dictionary identified with $[D]$.
- **Document:** x = ordered sequence of tokens $(x_1, \ldots, x_T)$, T = length of the document, $x_i \in [D]\ \forall i$
- **Vectorizer:** mapping φ transforming document x into vector $\varphi(x) \in \mathbb{R}^d$
- **Robust:** Hölder for **Hamming distance** on documents and **Euclidean distance** on embeddings:
$$\|\varphi(x) - \varphi(y)\| \leqslant L d_H(x,y)^\alpha .$$
- several classical choices for φ:
  - **TF-IDF:** term-frequency inverse document frequency, historical approach;
  - **concatenation:** associate each token to a vector u_e (*word vector*), add / concatenate positional embedding:
$$u(x_t, t) = [u_e(x_t); u_p(t)] .$$
    Then concatenate all word vectors together;
  - **ad hoc approaches:** learn an embedding from a dataset. This paper: doc2vec.

tokens	x_1	x_2	$\cdots$	x_T
token embeddings	$u_e(x_1)$	$u_e(x_2)$	$\cdots$	$u_e(x_T)$
	$+$	$+$		$+$
positional embeddings	$u_p(x_1)$	$u_p(x_2)$	$\cdots$	$u_p(x_T)$
embedding	$u(x_1)$;	$u(x_2)$;	$\cdots$;	$u(x_T)$

3: doc2vec embeddings (PVDM)

- **Key idea:** use document vector $q \in \mathbb{R}^d$ combined with local information to predict missing word in context
- **Context:** for a given window size ν,
$$\forall t \in [T], \qquad c(t) := (x_{t-\nu}, \ldots, x_{t-1}, x_{t+1}, \ldots, x_{t+\nu}) .$$
- **Local information:** average or concatenate one-hot vectors
$$\forall t \in [T], \qquad h_t := \frac{1}{2\nu} \sum_{s \in \gamma(t)} \mathbb{1}_{x_s} \in \mathbb{R}^D .$$
- **Parameters:** $P \in \mathbb{R}^{d\times D}$, $R \in \mathbb{R}^{D\times D}$
- **Prediction:** $\sigma(y_t)$, with
$$\forall t \in [T], \qquad y_t := R(Ph_t + q) = \pi_t + Rq \in \mathbb{R}^D .$$
- **Training:** on a corpus with N documents,
$$\underset{P,Q,R}{\text{Minimize}} \sum_{i=1}^N \frac{1}{T_i} \sum_{t \in x^{(i)}} -\log \sigma(y_t^{(i)})_{x_t^{(i)}} ,$$
- **Inference:** freeze P and R



4: Theoretical results

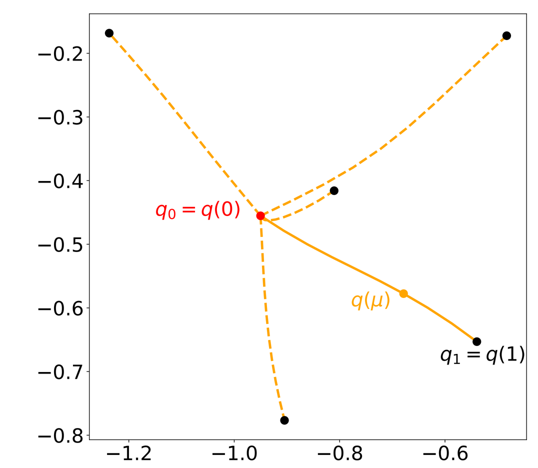
Theorem: concatenation, (TF-IDF,) and doc2vec embeddings are robust:
$$\|\varphi(x) - \varphi(y)\| \leqslant L d_H(x,y)^\alpha .$$

Concatenation	doc2vec
$L = \mathcal{O}\left(\max_{j,k \in \mathcal{S},\ j\neq k} \|u_e(j) - u_e(k)\|\right)$	$L = \mathcal{O}(1/T)$
$\alpha = 1/2$	$\alpha = 1$ (Lipschitz)

5: Proof ideas

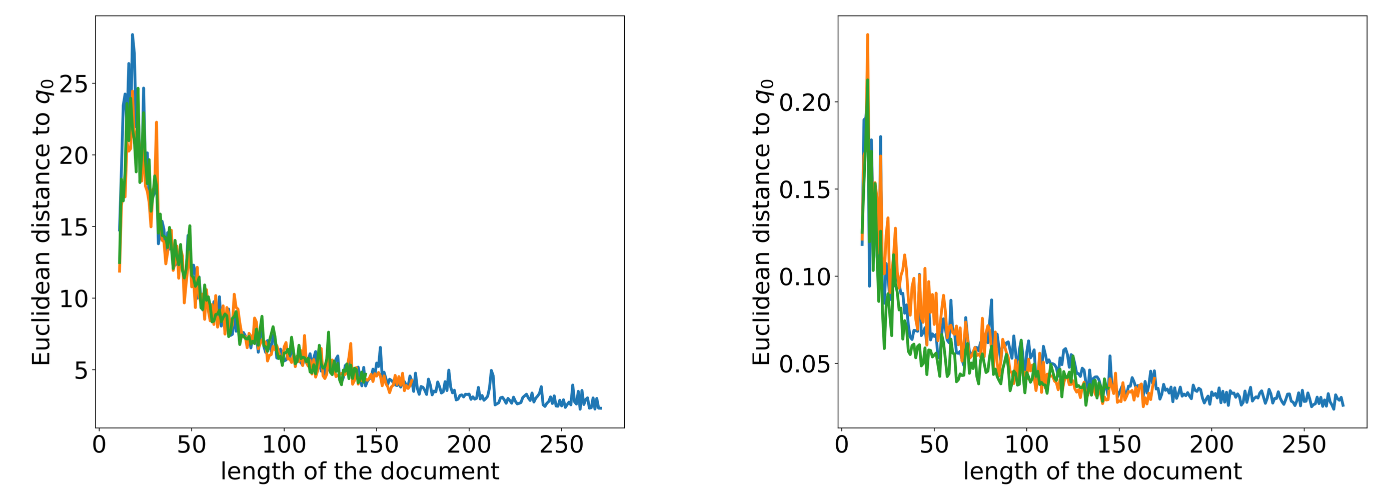
- **Key idea:** changing the document (x to $\tilde{x}$) changes the minimization problem
- we go from minimizing F to G, where
$$F(q) := \frac{1}{T} \sum_{t \in x} -\log(y_t)_{x_t} ,$$
  and G analogous for $\tilde{x}$
- we **interpolate** between F and G, minimizing
$$\forall \mu \in [0,1], \quad \Psi^{\mathrm{lin}}(\mu, q) := \mu G(g) + (1-\mu) F(q) .$$

- **fictitious embedding** $q(\mu)$
- dynamics of trajectory governed by an ODE
- precise control of this ODE, proving a **Grönwall–Bellman– Bahouri result**

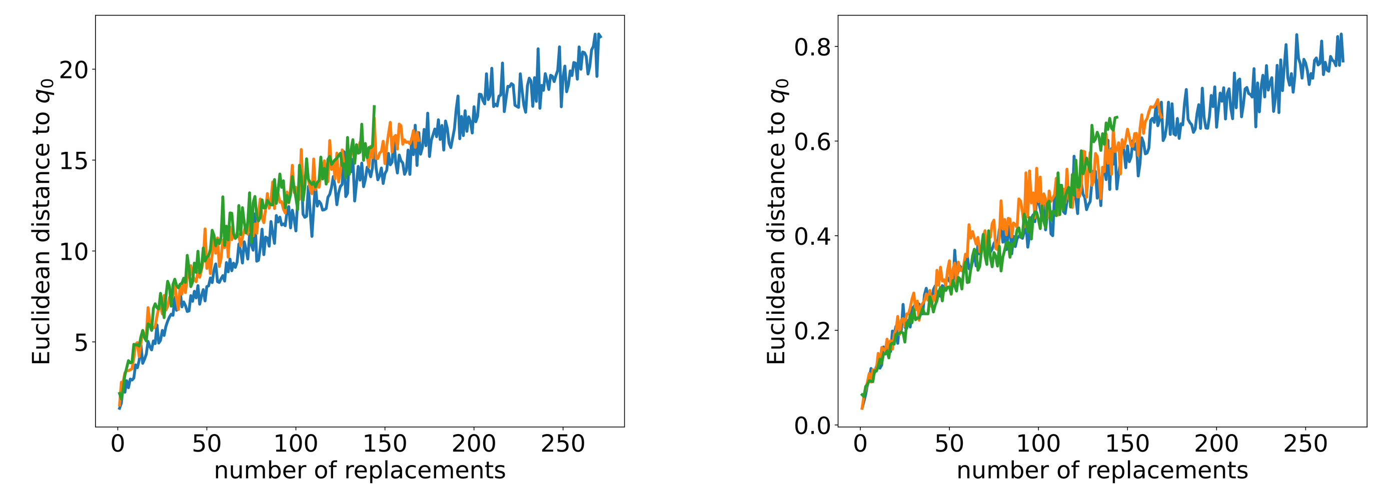


6: Experiments

Influence of the document length



Influence of the number of replacements



References

- Le, Mikolov, *Distributed representations of sentences and documents*, ICML, 2014
- Pachpatte, *On some new nonlinear retarded integral inequalities*, J. Inequal. Pure Appl. Math, 2004

correspondence to damien.garreau@univ-cotedazur.fr

This work was supported by ANR GraVa ANR-18-CE40-0005, NEMATIC ANR-21-CE45-0010, and NIM-ML ANR-21-CE23-0005-01