Embeddings and misspellings - Assessment and new methods

Gianluca Sperduti (ISTI-CNR)

N!ce t0 meet y0u!

Education and work experience:

- **Bachelor Degree:** Modern literature (Tor Vergata)
- **Master Degree:** Modern Philology (La Sapienza)
- **Specialization Course:** Big Data and Sociala Mining (University of Pisa)
- **Data Science Intern:** I worked as Data Science Intern in Freeda Media (4 months) - NLP projects
- **Junior Data Specialist:** I worked as a Junior Data Specialist in Vantea Smart (4 months)
- **ISTI-CNR fellowship &**
  National PhD in Artificial Intelligence: second (almost third) year (isti-cnr & University of Pisa)

Misspellings or we should say… mispeling?

an ungrammatical spelling of a word.

Msisepilnlgs

Human language is in constant evolution.

Alterations in a language, such as the use of non-standard language, the presence of grammatical and typographical errors, in any of its multiple declinations **do not pose a real problem for the cognitive capabilities of human readers.**

Misspellings can be...

Unintentional

QWERTY - *e.g. GPING*

SWAP - *e.g. GIONG*

ADD - *e.g. GOINNG*

REMOVE - *e.g. GONG*

And others...

Intentional (spelling variations)

GRAPHICAL - e.g. G01NG

SLANG - *e.g. GOIN'*

ADVERSARIAL ATTACKS -

Context

Recent state-of-the-art models, such as BERT, are not robust to **unseen** misspellings

Moreover, a common theoretical basis, a comprehensive survey, standard definitions, or shared frameworks, are still lacking.

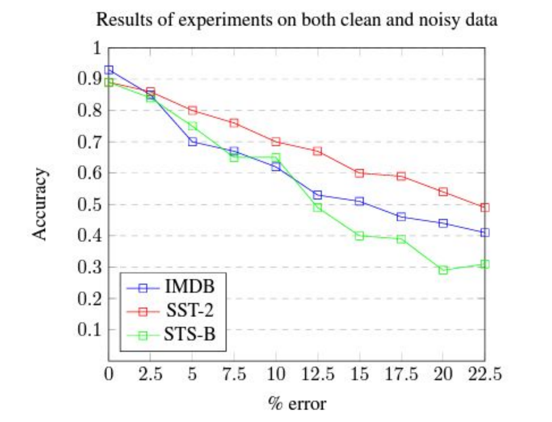


Figure 1: Accuracy vs Error

Kumar A, Makhija P, Gupta A (2020) Noisy text data: Achilles' heel of BERT. In: Proceedings of the 6th Workshop on Noisy User-generated Text (W-NUT@EMNLP 2020), Online, pp 16–21

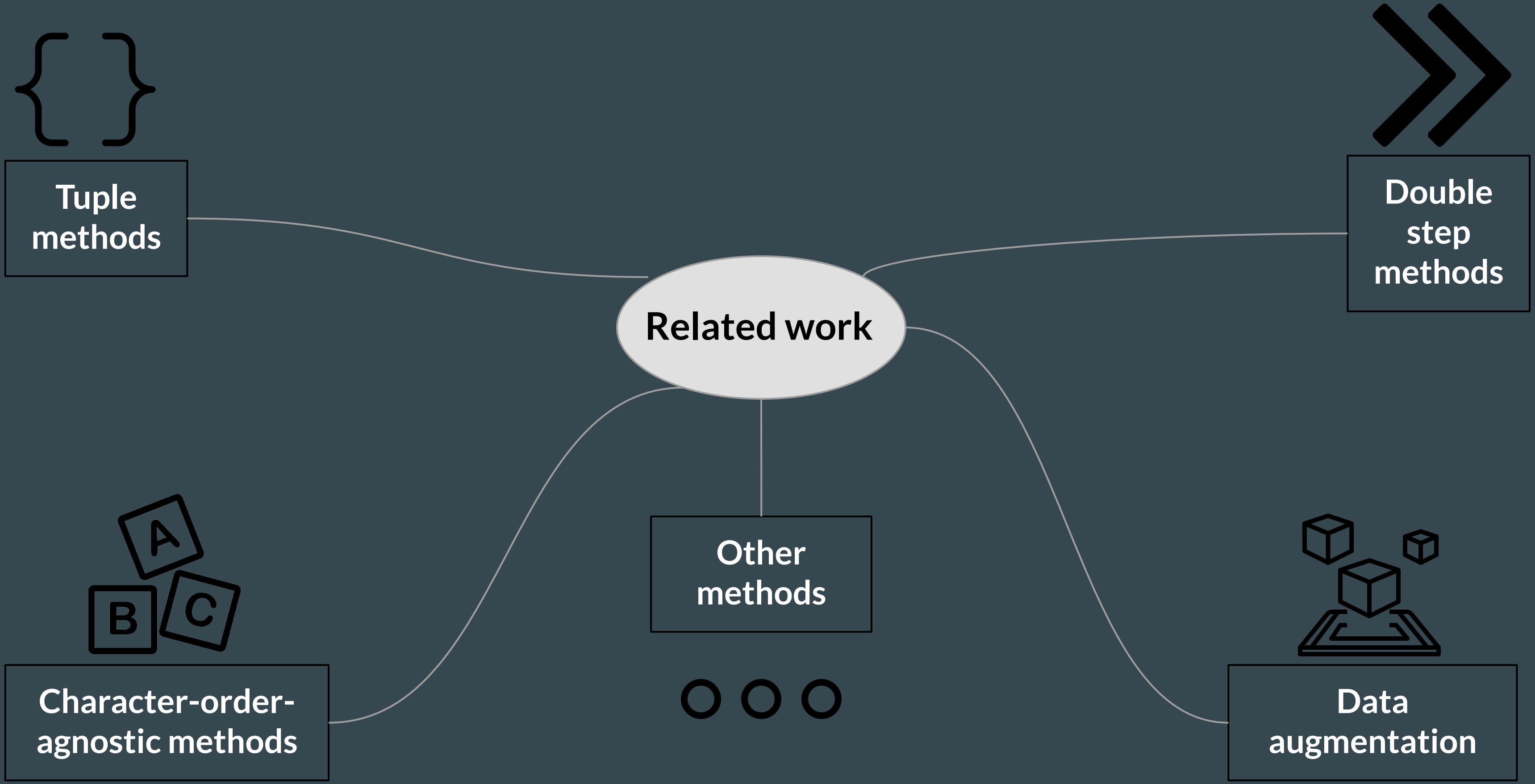
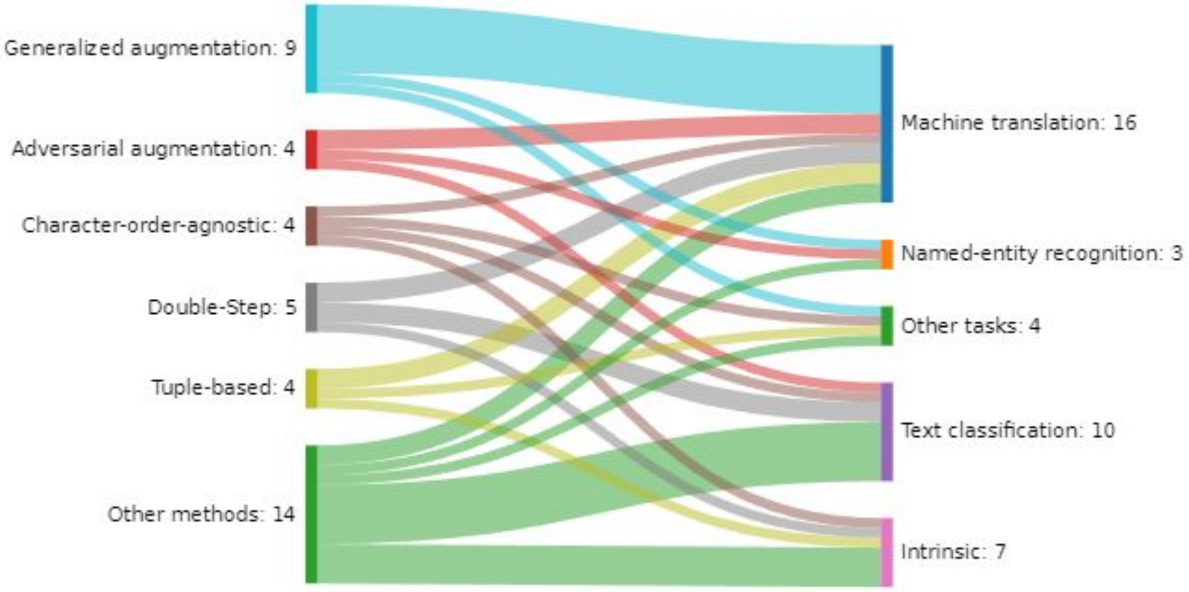

Tuple methods
Double step methods
Related work
Character-order-agnostic methods
Other methods
Data augmentation

Figure 5: Distribution of methods (left) across tasks (right)



Data augmentation

Represents one of the earliest attempts for solving the problem of misspellings in NLP tasks.

Classical data augmentation

Adversarial training

It also presents some **limitations** that are worth mentioning.

A costly solution...

Data augmentation entails an **additional cost** for modifying the training set, sometimes even resorting to complex techniques that seek to uncover the models' weakness.

...sometimes too static

Data augmentation typically **over-represent** certain types of misspellings, thus injecting sampling selection bias into the model.

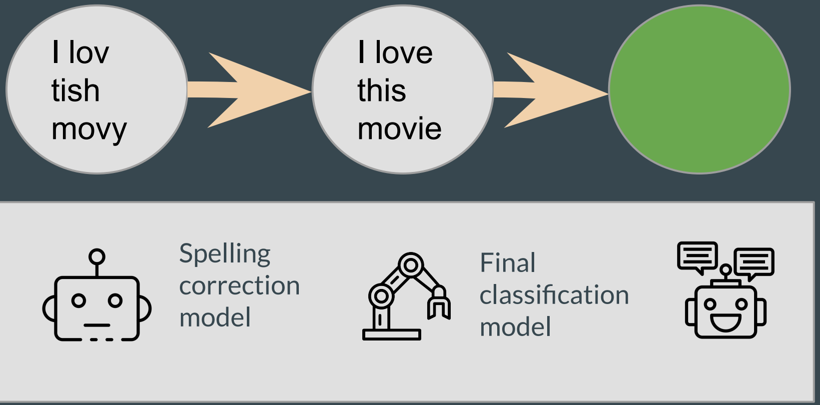
Double step with text normalization

The double step with text normalization method consists in using a two-step system to deal with misspellings.

It also presents some **limitations** that are worth mentioning.

Needs for two perfectly working models...

To get the best from this technique, we must employ two models: a spelling correction model and a downstream model. This is both costly and inefficient.



Character-order-agnostic method

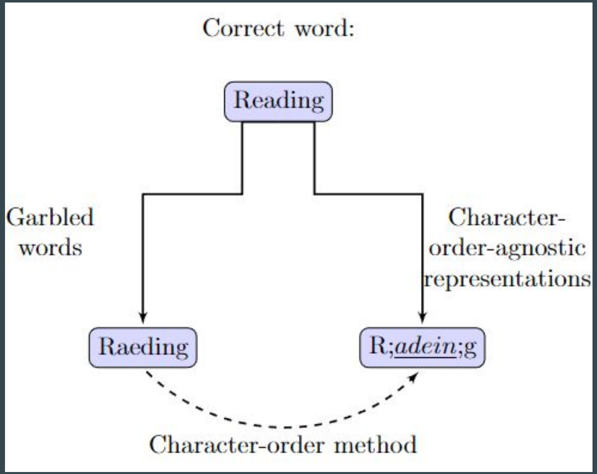
Character-order-agnostic methods gain inspiration from findings originating from the **psycho-linguistics** literature which indicate humans are able to read garbled words.

"reading"/"raeding"

It also presents some **limitations** that are worth mentioning.

A limited solution...

This approach targets one specific type of misspellings: Only misspellings based on confusion and exchange of internal letters can be tested with these ideas.



The tuple method

The tuple methods are highly heterogeneous, but share a common representation mechanism based on listing tuples of misspellings and correct spellings

{mispeling, misspelling}

Different approaches...

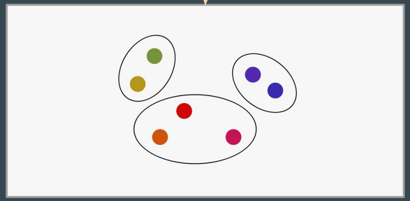
For example, Edizel et al. propose a FastText modification that changes the loss function including the distance between a correct and an incorrect word, while Zhou et al. gives a tuple of correct and incorrect sentences as input to a machine translation model.

Other methods

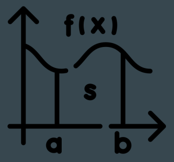
Methods that do not belong to any of the aforementioned groups, such as:

Cognitive-inspired methods

Experimental Encodings

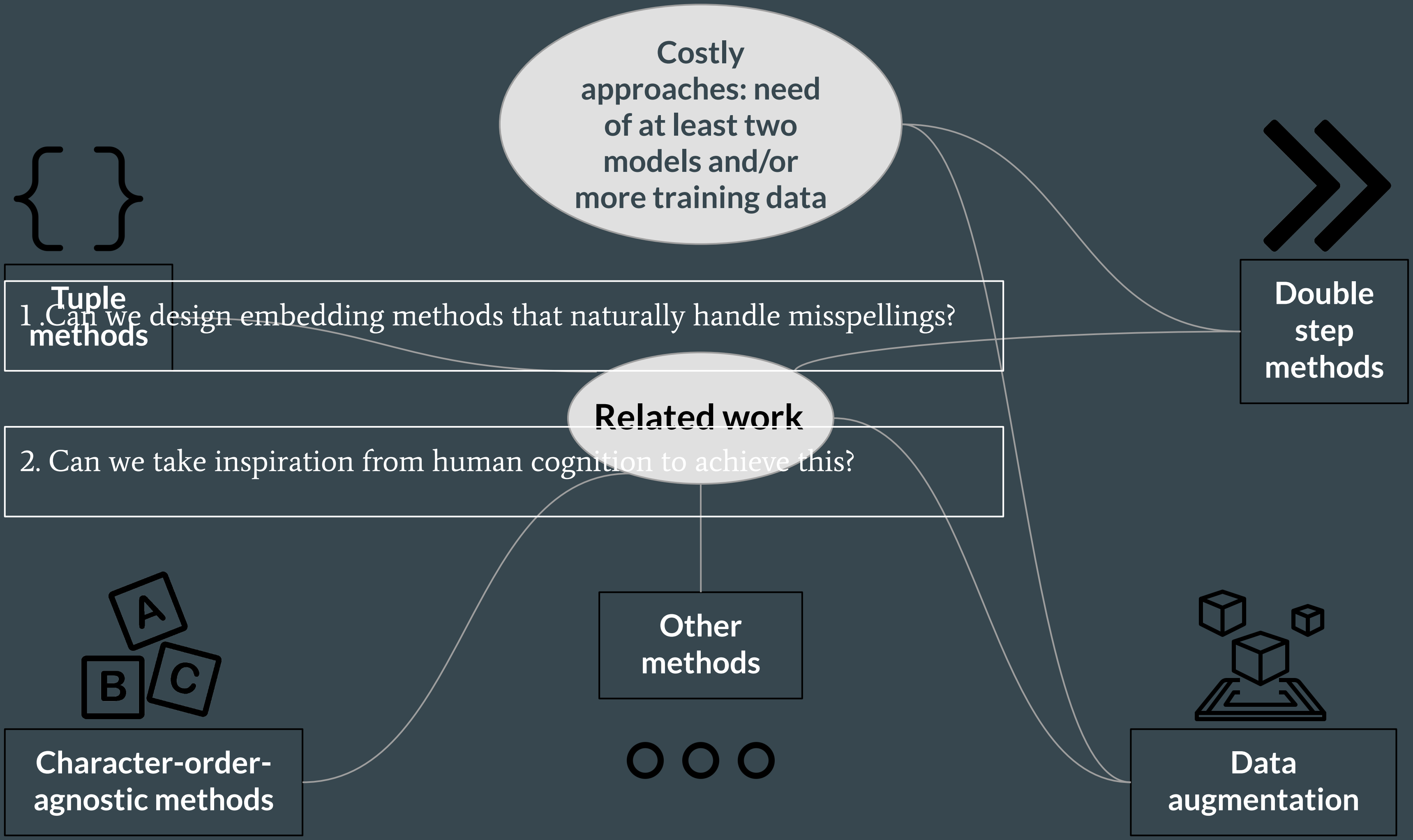


Regularization Functions



1 .Can we design embedding methods that naturally handle misspellings?

2. Can we take inspiration from human cognition to achieve this?



Some examples of techniques resilient to misspelling that we have tested ...

(1) Garbled-word embeddings

(2) Visually-grounded embeddings

“Aoccdrnig to a reasrech at Cmabrigde Uinervtisy, it deosn’t mttaer in waht oredr the ltteers in a wrod are, the olny itmopnrat tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe”

- Humans can read garbled text with low effort (backed by psycholinguistic studies).
- Can machines do something similar?
    - Previously reported experiments seem to indicate machines **cannot** do it...
- We believe machines **should be able too**.
    - We argue the key for achieving it comes down adopting order-invariant representations of text... and **it works**!

1

Garbled words -> Gralbed wrdos -> Gabelrd wdors

Garbled-word embeddings for jumbled text **(Best Short Paper Award, IIR, 2021).** We tested a character-order-agnostic method, called BE-sorting, on a benchmark of intrinsic tasks fo word embeddings using as training set the British National Corpus (trained in different versions, with the pre-processing technique, a garbling algorithm at various level of probability and with the normal version of the dataset). The method consists of a pre-processing operation, described as follows: Given a word $w = [c_1, c_2, \ldots, c_n]$ in which c_i denotes the character at position i, we sort alphabetically all the characters of each word, excluding the first and the last character (BE stands for: Begin, End).

Table 1

Performance evaluation of different sets of embeddings on 17 intrinsic-task benchmarks, grouped according to task (2nd row) and evaluation measure (3rd row).

	AP	BLESS	Battig	ESSLLI2c	ESSLLI2b	ESSLLI1a	MEN	WS353	WS353R	WS353S	MTurk	SL999	RW	RG65	Google	MSR	SE2012
	Categorization						Relatedness					Similarity			Analogy		
	Purity						Correlation					Correlation			Accuracy		
Garbled(0%)	.618	**.835**	**.376**	**.662**	**.765**	**.847**	.725	.635	.588	.682	**.553**	**.329**	**.160**	.782	**.262**	**.015**	.148
Garbled(5%)	**.640**	.818	.376	.653	.747	.836	**.728**	.635	.590	.683	.548	.324	.144	**.788**	.240	.012	**.153**
Garbled(10%)	.628	.819	.372	.640	.750	.822	.726	**.640**	.596	.680	.536	.320	.133	**.788**	.220	.009	.147
Garbled(50%)	.593	.804	.344	.600	.710	.795	.713	.627	**.606**	.662	.520	.267	.054	.735	.087	.002	.142
Garbled(100%)	.333	.539	.192	.566	.625	.663	.439	.253	.205	.288	.030	.140	.134	.377	.002	.000	.069
BE-sorted	.622	.833	.374	.644	.745	.841	.719	**.640**	.594	**.685**	.549	.324	.157	.785	.241	.015	.150
Full-sorted	.499	.626	.328	.515	.675	.659	.211	.249	.295	.250	.221	.124	.132	.049	.196	.009	.080
RandEmbeds	.159	.230	.092	.378	.525	.432	-.018	.127	.178	.048	-.074	.010	-.038	.006	.000	.000	.011

Submit Your Articles to the High-quality Academic Journals

From Elizabeth Jackson <elizabeth.has@joctr.net> on 2023-03-15 11:46

Details Plain text

If you wish to unsubscribe from such emails, you can click here.

Contribute Your Original Manuscript

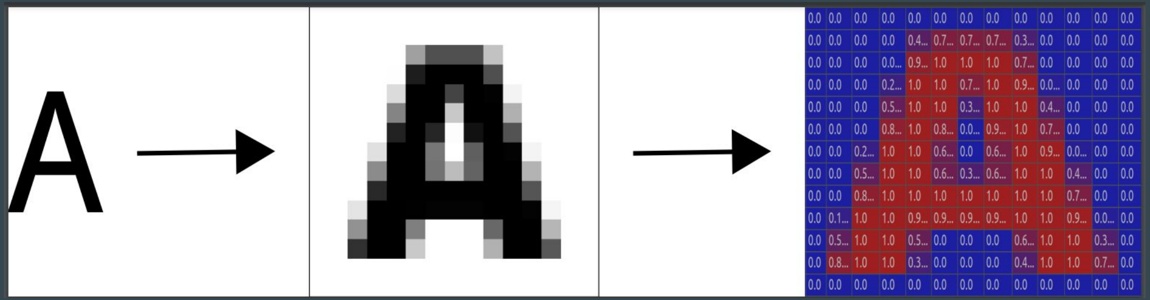
Dear Esuli, A; Moreo, A; Sebastiani, F; ...,

Warmest greetings from the assistant editor!

On our platform, there are over 300+ peer-reviewed academic journals covering nearly all academic fields. Our journals are created with the aim to offer good platforms for researchers and experts to share ideas and insights on all areas.

Become part of our Editorial Committee or Reviewer Team!

2 Visually-grounded embeddings



- **Visually-grounded embeddings:**
  - Initially proposed by Wang et al. [1].
  - Represent each character by rastering an image with a specific font and size
  - Embed the characters as a real-valued matrix representation from the pixels.

- **Methodological Differences:**
  - Wang et al.'s reduce the representation via PCA.
  - We do not reduce the dimension, but work directly with the original 10×16 images.
  - This allows us to generate visual representations **on the fly** for unseen characters.

[1] H. Wang, P. Zhang, E. P. Xing, Word shape matters: Robust machine translation with visual embedding, CoRR abs/2010.09997 (2020).

Homoglyphs:

- visually resemble each other.
- have different Unicode.
- can belong to non-Latin alphabets.
- can be generated by replacing letters with numbers.
- NLP model performance decreases (even char-based!)

4 Datasets:

- **(HS) Hate speech dataset from a white supremacy forum:** data from the Stormfront forum between 2002 and 2017.
- **(HATE) Automated hate speech detection and the problem of offensive language:** tweets.
- **(HASPEEDE):** Hate speech data from Italian Social Networks (specifically Italian Twitter and Italian Facebook).
- **(JIGSAW) The Kaggle's Jigsaw dataset (2017):** a collection of a large number of Wikipedia comments that have been labeled by human raters for toxic behavior.

4 Models:

- **SVM** based on sparse TFIDF vectors
- **CNN:** a convolutional char-based classifier operating on:
  - **CNN-R** random embeddings
  - **CNN-V** visually-grounded embeddings
  - **CNN-V_RL** visually-grounded embeddings processed by another CNN

3 Settings:

- **Clean:** no misspellings.
- **Adv:** injected in the 1,000 most important terms (weights from a linear model).
- **Hard_adv:** As above, but considering more difficult (i.e., different) chars

What we tested and did not work...

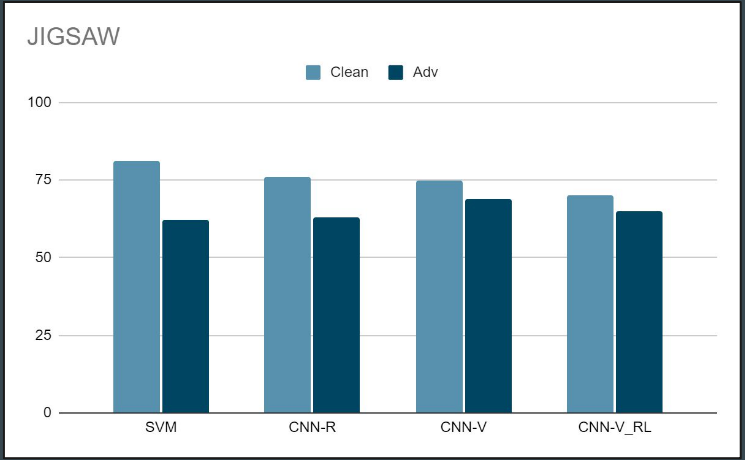
- We tested this character-based embeddings with Transformers, LSTM and RNN, but we could not train in a proper way: probably embeddings are too sparse to create a meaningful training with recurrent approaches.
- We also tested some dimensionality reduction techniques, with unsatisfying results.

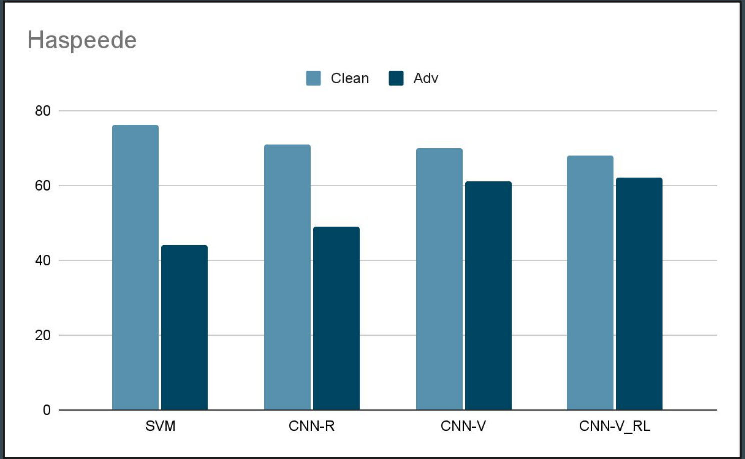
What are planning to do...

- A more comprehensive (final) test using also BERT, RoBERTa and AlBERTo (against our models), setting a wide range of “hardness levels”.
- Extends this idea into a language model (very preliminary idea).

JIGSAW (MACRO F1)	Clean	Adv	Hard_adv
SVM	**.816**	.628	-
CNN-R	.764	.632	.643
CNN-V	.752	**.699**	**.679**
CNN-V_RL	.705	.658	.646

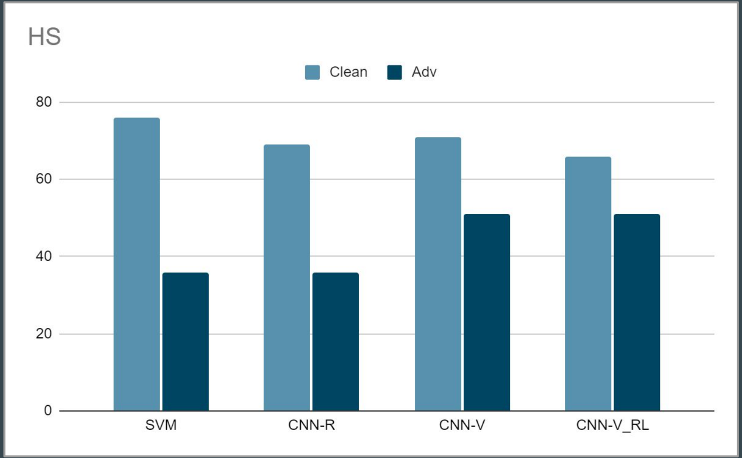
HASPEEDE (MACRO F1)	Clean	Adv	Hard_adv
SVM	**.76**	.44	-
CNN-R	.713	.491	.513
CNN-V	.707	.618	**.547**
CNN-V_RL	.682	**.629**	.546

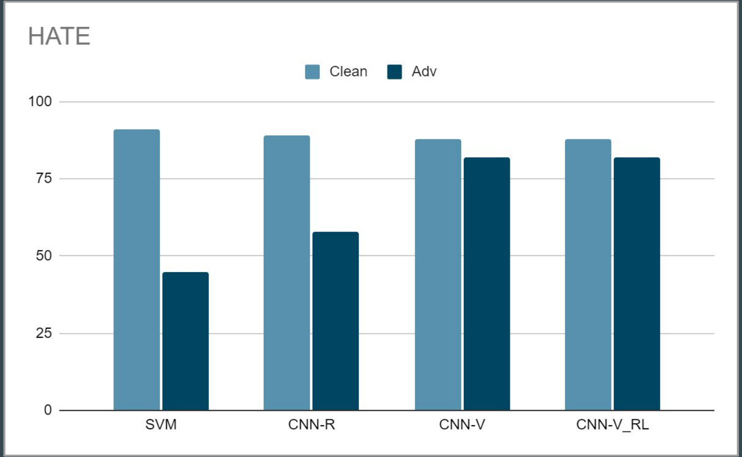




HS (MACRO F1)	Clean	Adv	Hard_adv
SVM	**.760**	**.366**	-
CNN-R	.699	**.366**	.361
CNN-V	.712	**.519**	**.456**
CNN-V_RL	.663	.518	.441

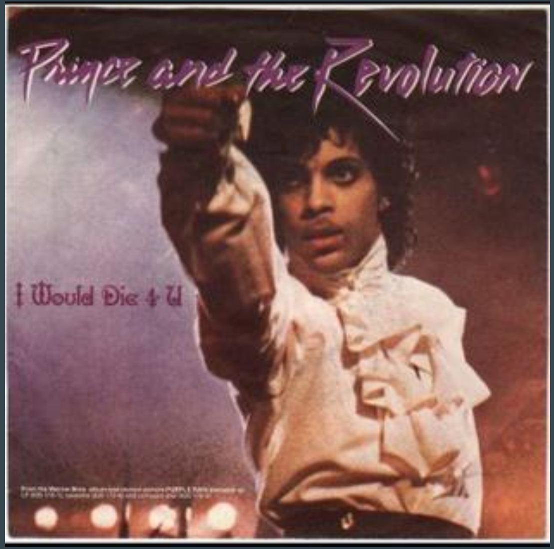
HATE (MACRO F1)	Clean	Adv	Hard_adv
SVM	**.91**	.458	-
CNN-R	.896	.583	.633
CNN-V	.887	**.829**	**.733**
CNN-V_RL	.883	.825	.722





Phonetic embeddings: can we use them to resist and capture "phonetic" spelling variations?

Phonetic spelling variations



Phonetic embeddings

Embeddings that represent language in a continuous vectorial space using phonetic characteristics of words/phonemes.

Closing up & Future Frontiers

- Need for standardized definitions and benchmarks.

- Misspellings resiliency should also provide resilience to language evolution diacronically, diatopically, and diastratically.

- Resilience to misspellings is crucial for the evolution of NLP systems.

- Models handling misspellings inspire more efficient representations.