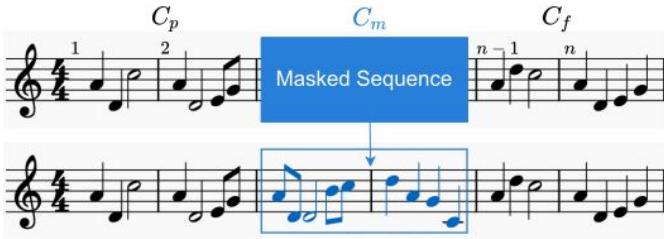


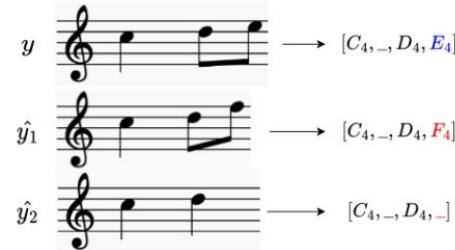
MUSIB: Musical Score Inpainting Benchmark

Mauricio Araneda-Hernandez^{1,3,4*}, Felipe Bravo-Marquez^{1,3,4}, Denis Parra^{2,3,4} and Rodrigo F. Cádiz^{4,5}

Task definition



Note Metrics



Current Issues:

- Non-compatible metrics across literature

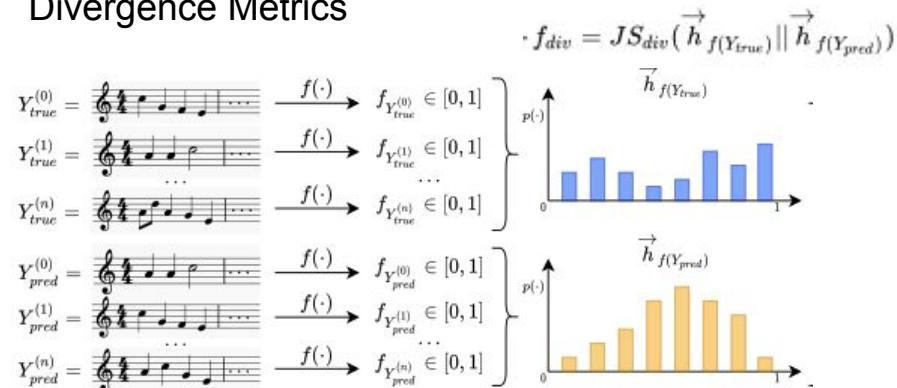
Main Goal:

- Standard for evaluation

Our proposal:

- Note Metrics
- Divergence Metrics

Divergence Metrics



SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA

José Cañete, Gabriel Chaperon, Rodrigo Fuentes
 Department of Computer Science, Universidad de Chile
 {jcanete, gchaperon, rfuentes}@dcc.uchile.cl

Jou-Hui Ho, Hojin Kang
 Department of Electrical Engineering, Universidad de Chile
 {jouhui.ho, ho.kang.k}@ug.uchile.cl

Jorge Pérez
 Department of Computer Science, Universidad de Chile &
 Millennium Institute for Foundational Research on Data (IMFD), Chile
 jperez@dcc.uchile.cl

ABSTRACT

The Spanish language is one of the top 5 spoken languages in the world. Nevertheless, finding resources to train or evaluate Spanish language models is not an easy task. In this paper we help bridge this gap by presenting a BERT-based language model pre-trained exclusively on Spanish data. As a second contribution, we also compiled several tasks specifically for the Spanish language in a single repository much in the spirit of the GLUE benchmark. By fine-tuning our pre-trained Spanish model we obtain better results compared to other BERT-based models pre-trained on multilingual corpora for most of the tasks, even achieving a new state-of-the-art on some of them. We have publicly released our model, the pre-training data and the compilation of the Spanish benchmarks.

ALBETO and DistilBETO: Lightweight Spanish Language Models

José Cañete^{1,3}, Sebastián Donoso^{1,3}, Felipe Bravo-Marquez^{1,3,4},

Andrés Carvallo^{2*}, Vladimír Araujo^{2,3,4,5}

¹ Department of Computer Science, University of Chile, Santiago, Chile

²Pontificia Universidad Católica de Chile, Santiago, Chile

³ National Center for Artificial Intelligence (CENIA), Santiago, Chile

⁴ Millennium Institute for Foundational Research on Data (IMFD), Santiago, Chile

⁵KU Leuven, Leuven, Belgium

jcanete@dcc.uchile.cl

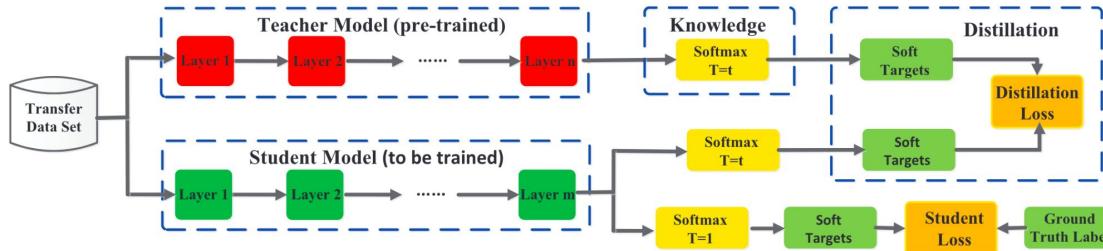
Abstract

In recent years there have been considerable advances in pre-trained language models, where non-English language versions have also been made available. Due to their increasing use, many lightweight versions of these models (with reduced parameters) have also been released to speed up training and inference times. However, versions of these lighter models (e.g., ALBERT, DistilBERT) for languages other than English are still scarce. In this paper we present ALBETO and DistilBETO, which are versions of ALBERT and DistilBERT pre-trained exclusively on Spanish corpora. We train several versions of ALBETO ranging from 5M to 223M parameters and one of DistilBETO with 67M parameters. We evaluate our models in the GLUEs benchmark that includes various natural language understanding tasks in Spanish. The results show that our lightweight models achieve competitive results to those of BETO (Spanish-BERT) despite having fewer parameters. More specifically, our larger ALBETO model outperforms all other models on the MLDoc, PAWS-X, XNLI, MLQA, SQAC and XQuAD datasets. However, BETO remains unbeaten for POS and NER. As a further contribution, all models are publicly available to the community for future research.

Speedy Gonzales: A Collection of Fast Task-Specific Models for Spanish

José Cañete
 Universidad de Chile
 Santiago, Chile
 jose.canete@ug.uchile.cl

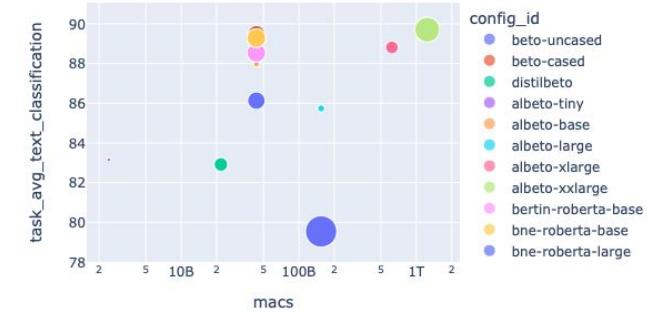
Felipe Bravo-Marquez
 Universidad de Chile
 Santiago, Chile
 fbravo@dcc.uchile.cl



Spanish Language Models

José Cañete

<https://josecannete.github.io/>
josecannete@gmail.com



También pueden preguntarme por **modelos multimodales vision-language**: CLIP / Multimodal Transformers / etc

RiverText: A Python Library for Training and Evaluating Incremental Word Embeddings from Text Data Streams

Gabriel Iturra-Bocaz

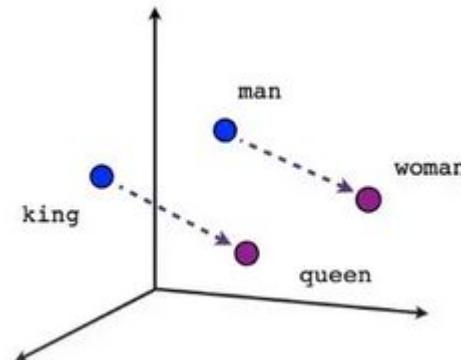
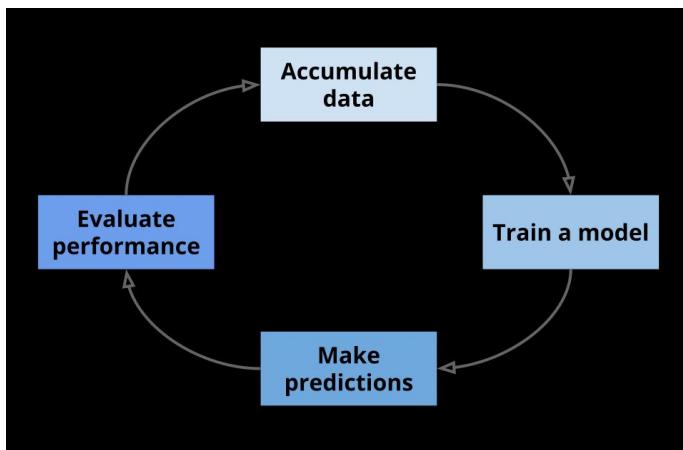
gabrieliturrab@ug.uchile.cl

Department of Computer Science, University of Chile
Santiago de Chile, Chile

Felipe Bravo-Marquez

fbravo@dcc.uchile.cl

Department of Computer Science, University of Chile
Santiago de Chile, Chile



Incremental Word Embedding

Gabriel Iturra

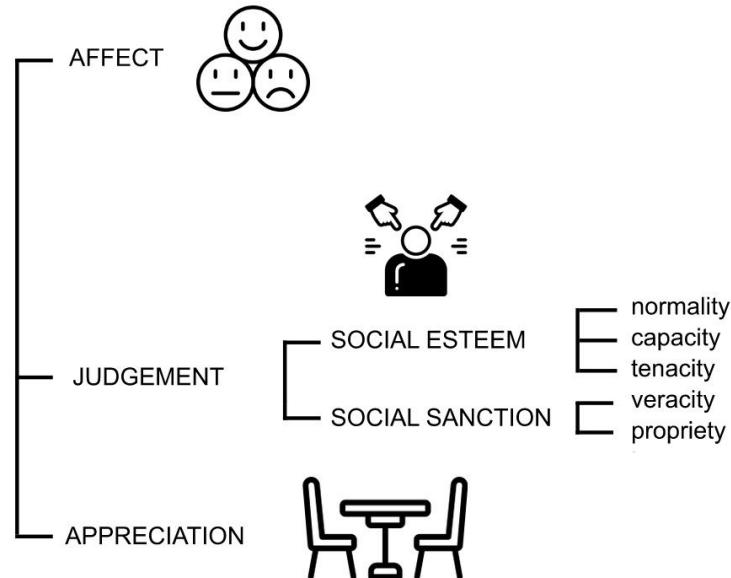
<https://giturra.cl/>
gabrieliturrab@ug.uchile.cl

Attitude Analysis: A linguistic-based task for detailed position detection in texts

Jorge Ortiz Fuentes
jorge@ortizfuentes.com



ATTITUDE



Multilingual and Multi-Domain Crisis Classification in Social Media

Cinthia Sánchez - users.dcc.uchile.cl/~cisanche

Crisis Informatics challenges

- Most approaches are monolingual (mainly focused on English).
- For many languages, there are limited or no labeled data.
- Information relevance may vary across responder role, domain, and other factors.

Mensajes	¿Relevante para la crisis?
Ascienden 52 los muertos y 1,2 millones de afectados por terremoto en Guatemala: La cifra de muertos por el... http://t.co/YyCHBarU	Si
No sé si preocuparme porque no sentí el temblor de ahorita o qué.	No
...	

Approaches for addressing the multilingual data scarcity

Transfer learning between languages and domains for detecting crisis-related messages in low-resource scenarios.

[Accepted paper in International AAAI Conf. on Web and Social Media, ICWSM 2023]



Applying this knowledge transfer learning approach to other critical tasks on multilingual messages, e.g., categorization of humanitarian information.

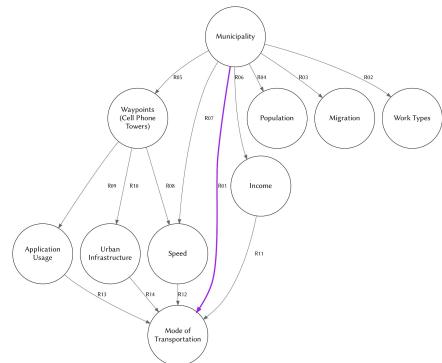
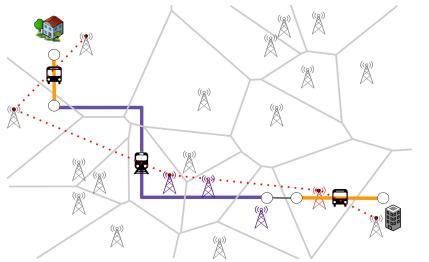


Multi-task learning to leverage the multilingual resources of multiple tasks.

Feel Old Yet?

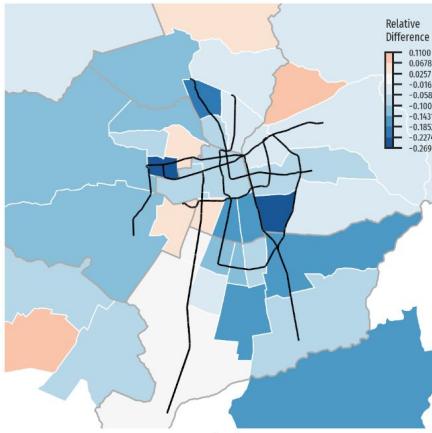
Updating Mode of Transportation Distributions from Travel Surveys using Data Fusion with Mobile Phone Data

Eduardo Graells-Garrido (DCC U. de Chile, CENIA)
Daniela Opitz (U. del Desarrollo)
Francisco Rowe (U. of Liverpool)
Jacqueline Arriagada (U. of Leeds)

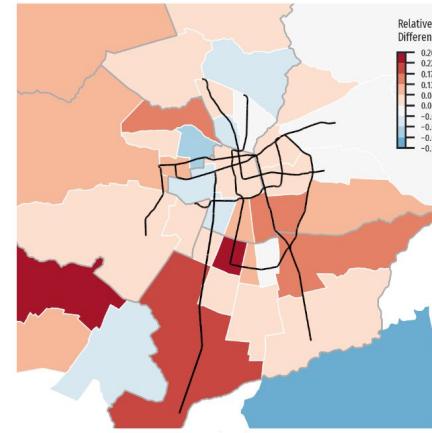


Differences in Modal Share (from 2012 to 2020)

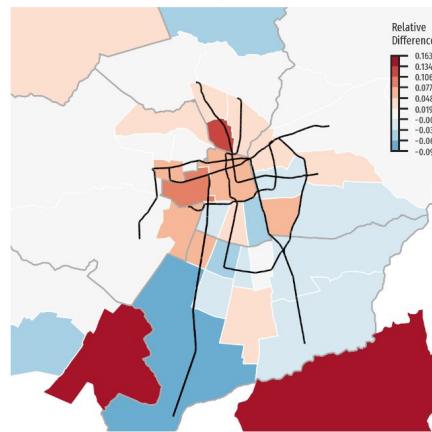
a) Mass Transit



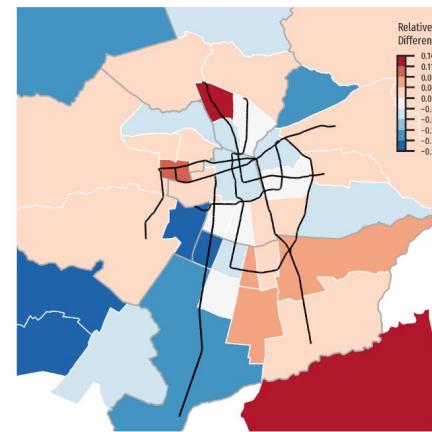
b) Motorised



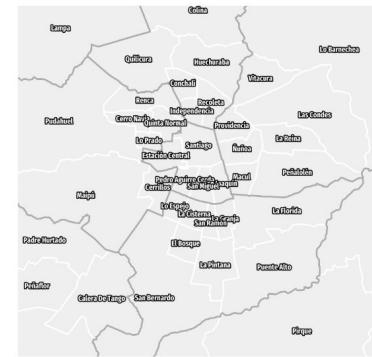
c) Taxi



d) Active



Municipalities



Macro Areas



Lexical Semantic Change in Spanish

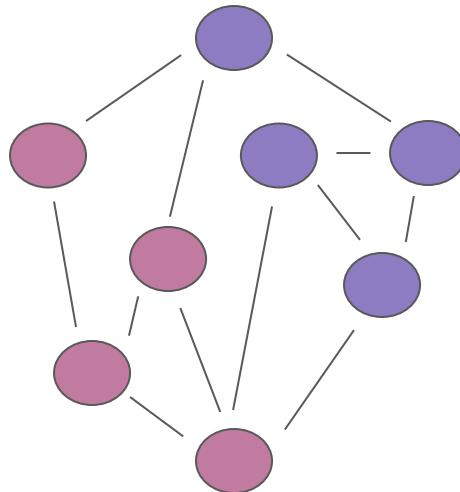
Frank D. Zamora Reina
fzamora@dcc.uchile.cl

Motivation

mouse ----- mouse
(animal) (device)

Applications

- . key to models of language and the cultural evolution of society
- . applicable to solve real-world historical semantics/lexicography problems
- . diachronic text evaluation



Word Usage Graph

Structure

- . diachronic text
- . annotations
- . models

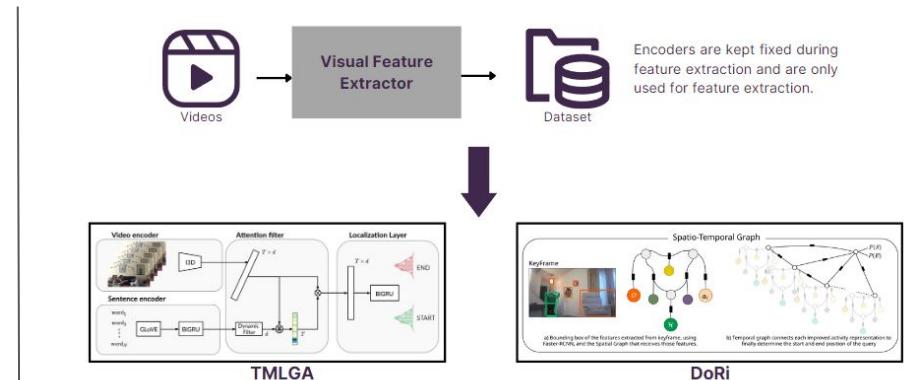
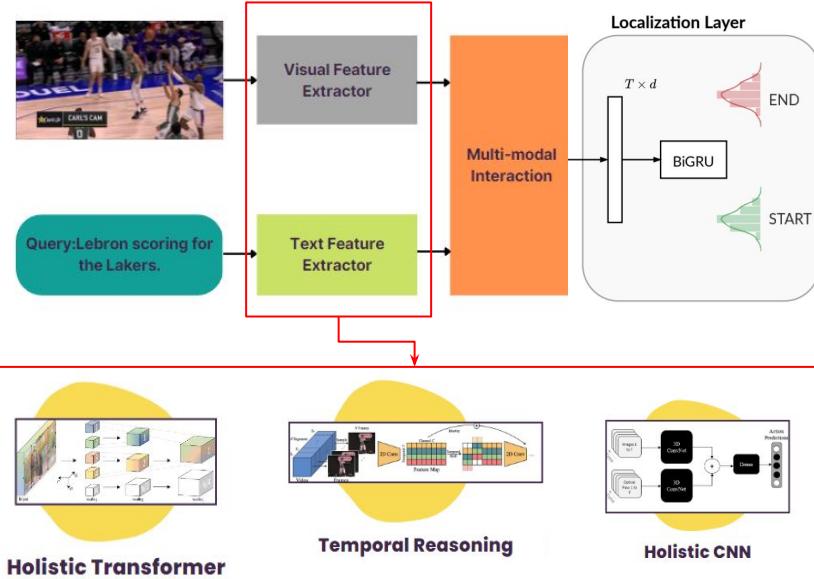
Approaches

- . contextualized methods, intermediate representations and change measure
- . proxy tasks (WSD, WSI, WiC)
- . currently, we're applying clustering methods on graphs

Benchmarking video action features for the video temporary sentence grounding task

Ignacio Meza
imezadelajara@gmail.com

Given an untrimmed video, temporal sentence grounding in videos is to retrieve a video segment, also known as a temporal moment, that semantically corresponds to a query in natural language.



Goal:

Studying the impact of different video classifiers on the task of temporal localization of a relevant moment.

Contribution:

- Generate a feature dataset for the tgsv task.
- Observe the impact of encoders of different nature.

Felipe Bravo-Marquez

Research with Students

- Semantic Change Detection
- Fairness in Word Embeddings
- Incremental Word Embeddings
- Framing and Polarization in Social Media
- Other.....

Open Source Software Projects

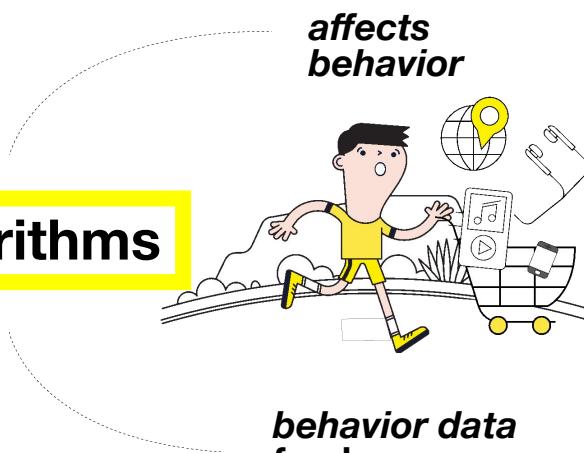
- WEFE
- DashAI
- RiverText
- Spanish Language Models

Study Time:

- Causal Inference
- Reinforcement Learning

Algorithmic Recommendations

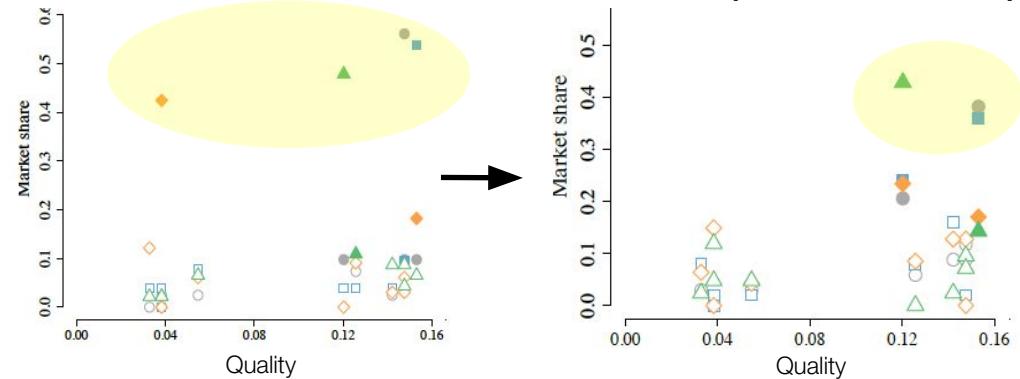
Andrés Abeliuk, DCC, CENIA



Algorithms

Ranking Algorithms

[Abeliuk et al., WWW 2017]



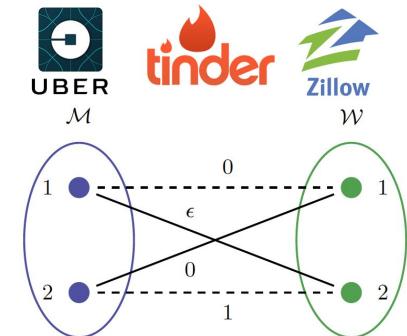
Matching Algorithms

Maximize social welfare

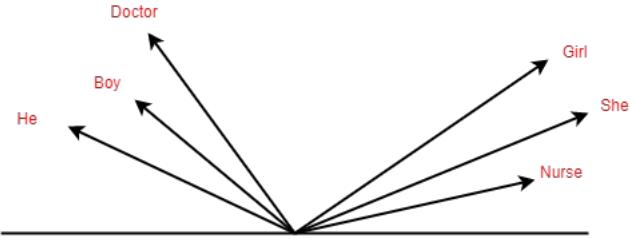
vs

Maximize retention

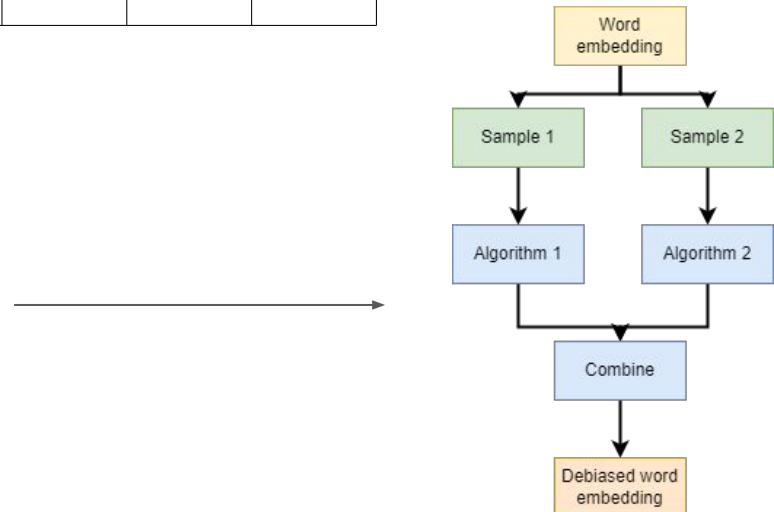
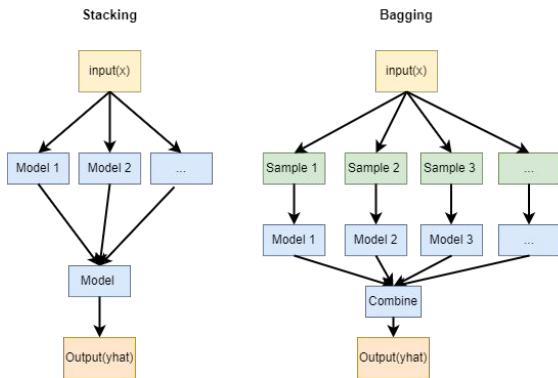
[Abeliuk et al., TEAC 202?]



Bias Mitigation Algorithms for Word Embeddings



Algorithms				
Normalization	HD	DHD	HSR	RAN
Word Sets	Def. pairs + Bias Definition	Def. pairs + Bias Definition	X	X
			Bias Definition	Def. pairs + Bias Definition

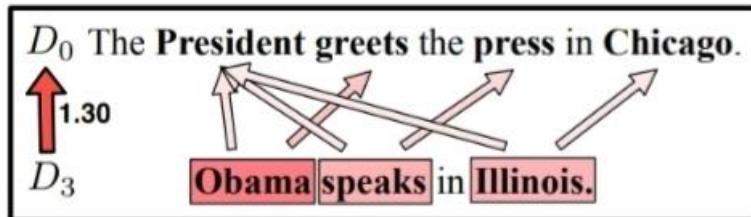
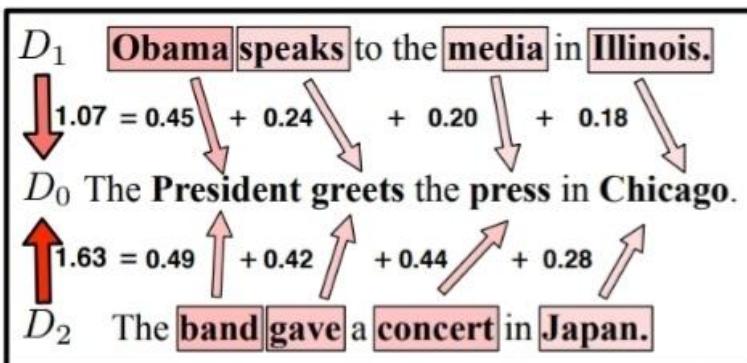


Evolucion de temas en noticias nacionales

Interpretabilidad en la distancia de documentos

Clemente Henriquez

From Word Embeddings To Document Distances



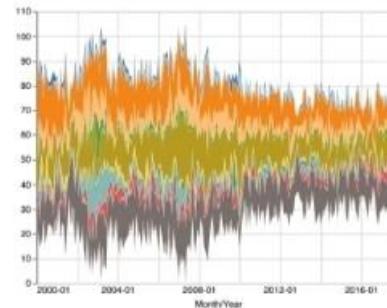
A Systematic Media Frame Analysis of 1.5 Million New York Times Articles from 2000 to 2017

[Please cite the WebSci'20 version of this paper]

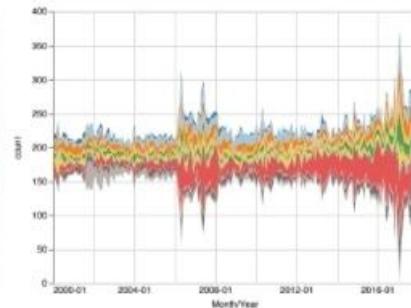
Haewoon Kwak
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
haewoon@acm.org

Jisun An
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
jisun.an@acm.org

Yong-Yeol Ahn
Indiana University, Bloomington
IN, USA
yyahn@iu.edu



(c) Smoking



(d) Immigration

<https://www.kimalmapuzugun.cl/>

I'm emailing to see if you were interested in trying to organize a workshop on Natural Language Processing for the Indigenous peoples of Chile (including Mapuzugun :). We could potentially get funding from here:

https://naacl.org/calls/emerging_regions/

Perhaps we could do a week-long NLP workshop that combines:

- fast-track or crash-course lectures on NLP and potentially linguistic concepts
- a hackathon where CS students (and ideally also representatives of the Mapuche community) -- or even better, Mapuche CS students 😊
- work on building tools/datasets/etc for Mapuzugun

Thoughts?

[https://naacl.org/calls/
emerging_regions/](https://naacl.org/calls/emerging_regions/)

Uchile Relela lightning talk

Valentin Barriere

18/01/23

Who Am I?

I am a researcher in NLP, got my PhD in 2019 about "*Opinion Analysis in Speech Interactions*". I was previously working for the EC's Joint Research Center

My interests:

- Multimodality
- Interactions dynamics
- Multilinguality
- Bias detection in sentiment analysis



Empathy, Distress and Emotion

- Nely: a social robot with **empathy**



- **Emotion** in disaster response



- **Personality** detection and adaptation for a virtual agent



We organize the WASSA workshop: Computational Approaches to Subjectivity, Sentiment & Social Media Analysis

- Held at ACL 2023 in Toronto
- Shared task on **Empathy, Distress and Emotion Detection in Conversations**



Last year WASSA:

<https://wassa-workshop.github.io/2022/>

Participatory Democracy

- Participatory Democracy
- Large scale implies automation
- Europe has high diversity of languages, topics, propositions and arguments



Participatory Democracy – Example

Title	Topic	Proposal	Comment	Stance	url
Focus on Anti-Aging and Longevity research	Health	The EU has presented their green paper on ageing, and correctly named the aging...	The idea of prevention being better than a cure is nothing new or revolutionary. Rejuvenation...	Pro	
Set up a program for returnable food packaging...	Climate change and the environment	The European Union could set up a program for returnable food packaging made from...	Bringing our own packaging to stores could also be a very good option. People would be...	Pro	
Impose an IQ or arithmetic-logic test to immigrants	Migration	We should impose an IQ test or at least several cognitive tests making sure immigrants have...	On ne peut pas trier les migrants par un simple score sur les capacités cognitives. Certains fuient la guerre et vous...	Against	
Un Président de la Commission directement élu...	European democracy	Les élections, qu'elles soient présidentielles ou législatives, sont au coeur du processus...	I prefer sticking with a representative system and have the President of the...	Against	
Europa sí, pero no así	Values and rights, rule of law, security	En los últimos años, las naciones que forman parte de la UE han visto como su soberanía ha sido...	Zdecydowanie nie zgadzam się z pomysłem, aby interesy indywidualnych Państw miały...	Against	

Table 1: Examples of comments and proposals with the associated stance

Participatory Democracy – Challenge @CLEF

- The CF_S and CF_U datasets will be available on demand
- CF_E will be used as a test set of one of the Touché Lab @ CLEF23 in Thessaloniki: **Intra-Multilingual Multi-Target Stance Classification**



- CF_S will be used as data for the ValueEval task SemEval 2023: **Identification of Human Values behind Arguments**



Prompt-based data-augmentation

Since CLIP, new models allow to generate and modify images with text.

These models can generate situations that are strange

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup
riding a horse lounging in a tropical resort in space playing basketball with cats in space
in a photorealistic style in the style of Andy Warhol as a pencil drawing

DALL-E 2



Prompt-based data-augmentation

Since CLIP, new models allow to generate and modify images with text.

These models can generate situations that are strange, funny

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of

soup

that is a portal to another
dimension **that looks like a**
monster as a planet in the universe

knitted out of wool spray-painted on a
wall made out of plasticine

DALL-E 2



Prompt-based data-augmentation

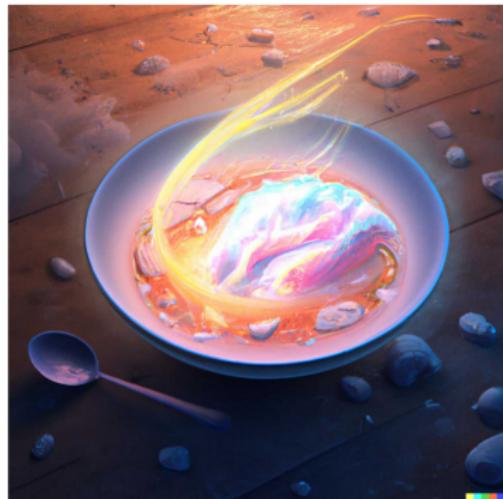
Since CLIP, new models allow to generate and modify images with text.

These models can generate situations that are strange, funny or artistic.

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup
that is a portal to another dimension that looks like a monster as a planet in the universe
as digital art in the style of Basquiat drawn on a cave wall

DALL-E 2



Prompt-based data-augmentation

Let's take an example:

- Detect attribute about gender, color, emotion, etc... in a caption to:
 - Evaluate the performance of a model
 - Modify the captions and generate images for data-augmentation purposes



"A girl with a {COLOR} scarf and ear phones":



"A {GENDER} with a black scarf and ear phones":

Questions?

References i
