Methodological Issues in Using Word Embeddings in a Sociolinguistic Perspective: The Case of Contact-Induced Semantic Variation Across Canadian Twitter Corpora

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

To cite this version:
Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy. Methodological Issues in Using Word Embeddings in a Sociolinguistic Perspective: The Case of Contact-Induced Semantic Variation Across Canadian Twitter Corpora. Empirical Studies of Word Sense Divergences across Language Varieties, Mar 2020, Hamburg, Germany. hal-02502916

HAL Id: hal-02502916
https://hal.archives-ouvertes.fr/hal-02502916
Submitted on 9 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Methodological issues in using word embeddings in a sociolinguistic perspective: the case of contact-induced semantic variation across Canadian Twitter corpora

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy
CLLE, CNRS & University of Toulouse (France)
(filip.miletic, anne.przewozny, ludovic.tanguy)@univ-tlse2.fr

Introduction

**SEMANTIC VARIATION IN QUEBEC ENGLISH**
- Quebec English (QcE) is a regional variety of Canadian English spoken by a minority of Quebecers.
- A possible consequence of contact with Quebec French: English words used with meanings typical of French cognates.
- Existing descriptions (Fee, 1991; 2008; Boberg, 2012; Rouaud, 2019) do not explore the extent or the precise status of this phenomenon.

**AN INTERDISCIPLINARY APPROACH**
- Aim: systematically identify semantic variants specific to QcE and investigate speaker-level factors driving this variation.
- Data collection and analysis draw on variationist sociolinguistics (Labov, 1972; Tagliamonte, 2012).
- Word embeddings are used to computationally detect synchronic semantic variation (e.g. Del Tredici & Fernández, 2017).

Methodology

**DATA**
- Corpus of tweets published from 2015 onward
- Search API to identify users ⇒ timeline crawl
- User-level filtering: location, language, near-duplicate exclusion

**ANALYSIS**
- A word embedding model was trained for each subcorpus using word2vec (Mikolov et al., 2013).
- Models were aligned using Orthogonal Procrustes as in diachronic studies (Hamilton et al., 2016).
- Prominent divergences in Montreal were detected using cosine-distances between a word’s vectors.

**RESULTS**
- Differences in spellings (e.g. in Montreal vs. Vancouver)
- Prominent divergences in Montreal reflect contact-induced semantic variation.
- The variation we study is prominent in Montreal due to a single Twitter account.

**Methodology issues**

**RESULTS OF LIMITED INTEREST**
- Cultural factors: unsupervised refers to machine learning due to Montreal’s IT industry; *chum* denotes a species of salmon in Vancouver due to its geography.
- Local referents: plateau refers to the borough of Plateau-Mont-Royal in Montreal.
- Prolific users: waffle ‘make waffles’ (rather than ‘be undecided’ or ‘speak at length’)

**GENERAL ISSUES**
- The variation we study is subtler than long-term semantic change: conventional and contact-related meanings coexist ⇒ polysemy.
- Regional regularities do not suffice to explain this variation: we need to identify and describe speakers with similar behaviors.
- Uncertainty over the relationship between computational results and real-life sociolinguistic behaviors.

Ongoing work

**PREPROCESSING**
- Word-level language ID
- Topic modelling

**WORD EMBEDDING MODELS**
- Contextual word embeddings ⇒ polysemy, user clusters

**SOCIOLINGUISTIC FIELDWORK**
- Cohort study based on a group of native speakers reflecting linguistic profiles from the Twitter corpus.
- Aim: examine the status and representations of the variants detected using word embeddings.
- The results will inform future studies in computational sociolinguistics.

Conclusions

- This study has brought to light important descriptive observations, particularly the role of bilingualism and the importance of polysemy in contact-induced semantic variation.
- Methodological issues obfuscate relevant results in our models, but they are being addressed through ongoing work.
- The need for our interdisciplinary approach is already clear: word embedding models are needed to detect semantic variation, and fine-grained sociolinguistic analysis to clarify its nature.

References


Key examples

**A PREVIOUSLY DESCRIBED CASE: exposition**

I really want to go to an art museum or an art exposition. Canada’s centennial year saw Montreal host the 1967 International and Universal Exposition. On parle de notre exposition Brown’s Hill! // An article about our exposition Brown’s Hill

Three straight scenes of clunky dialogue filling in for exposition. Yup, it’s a Schwarzenegger film! A brilliant exposition of dietary fiber & the wonders it can perform for human health.

**A NEWLY IDENTIFIED CASE: definitively**

Pouring coffee beans in the water tank... I’m definitively need coffee!! That’s most definitely a 10 Im getting tattooed right now and it’s definitively the one that hurt the most.

Worse, the research community has performed multiple trials & studies that all definitively show no connection between vaccination & autism

I think we can definitively say Carey Price is permanently broken.

- Incidents of autism
- Literature review on autism and vaccination
- Methodology for identifying cases
- Conclusions on causality

- Impact on public perception
- Policy implications

- Further research on autism and vaccination

- Improved methodologies for analysis