Extracting dialect-specific features from dialect classifiers

Yves Scherrer and Dana Roemling
(Joint work with Aleksandra Miletić and Noëmi Aepli)

ICLAVE 12, Vienna
Panel *Embracing Variability in Natural Language Processing*
10 July 2024

How?

1. Make trained models more robust to variability
   - Aepli & Sennrich (2020): Simulate variation as character-level noise
2. Make evaluation measures both more robust and more sensitive to variability
   - Aepli et al. (2023): Adapt COMET to Swiss German
3. Inform the model about varieties in a multi-task setup
   - Scherrer & Rabus (2019): Use variety as an additional feature in morphological tagging
4. Analyze variety representations inferred during training
   - Kuparinen & Scherrer (2023): Speaker label embeddings in normalization systems represent dialects
5. Determine the features (e.g., words) that are characteristic for certain dialects

Determine dialect-specific features

The dialectological point of view:

- Dialectometric methods (clustering, dimensionality reduction) provide a (relatively) objective, high-level classification of dialects
- Tracing back such classifications to individual features (e.g. words) is an important desideratum of dialectologists
- But it's challenging...
    - Prokić et al. (2012): *Detecting shibboleths*
    - Rumpf et al. (2009, 2010): Factor analysis
    - Topic modelling approaches: Eisenstein et al. (2010), Hovy & Purschke (2018), Kuparinen & Scherrer (2024)

Determine dialect-specific features

The NLP point of view:

- Relatively easy to understand traditional machine learning models and reconstruct their decision processes
- Not true anymore for neural-network-based models, and even less so for pre-trained models
- A growing list of works that focus on **interpretability and explainability** of NN-based models
  - Intrinsic approach: Add some additional mechanisms to the neural network and train them jointly with the rest of the model
  - Post-hoc/extrinsic approach: Obtain insights from the predictions of an existing, unmodified model
  - Application to dialect identification: Xie et al. (2024)

Interpretable dialect classifiers

Extracting Lexical Features from Dialects via Interpretable Dialect Classifiers

Roy Xie♣♦ **Orevaoghene Ahia**♠ **Yulia Tsvetkov**♠ **Antonios Anastasopoulos**♦

♣ Duke University

♠ Paul G. Allen School of Computer Science & Engineering, University of Washington

♦ Department of Computer Science, George Mason University

ruoyu.xie@duke.edu {oahia, yuliats}@cs.washington.edu antonis@gmu.edu

Abstract

Identifying linguistic differences between dialects of a language often requires expert knowledge and meticulous human analysis. This is largely due to the complexity and nuance involved in studying various dialects. We present a novel approach to extract distinguishing lexical features of dialects by utilizing interpretable dialect classifiers, even in the absence of human experts. We explore both post-hoc and intrinsic approaches to interpretability, conduct experiments on Mandarin, Italian, and Low Saxon, and experimentally demonstrate that our method successfully identifies key language-specific lexical features that contribute to dialectal variations.[1]

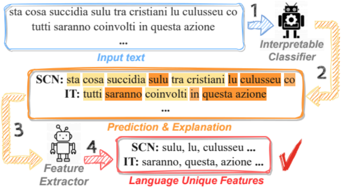
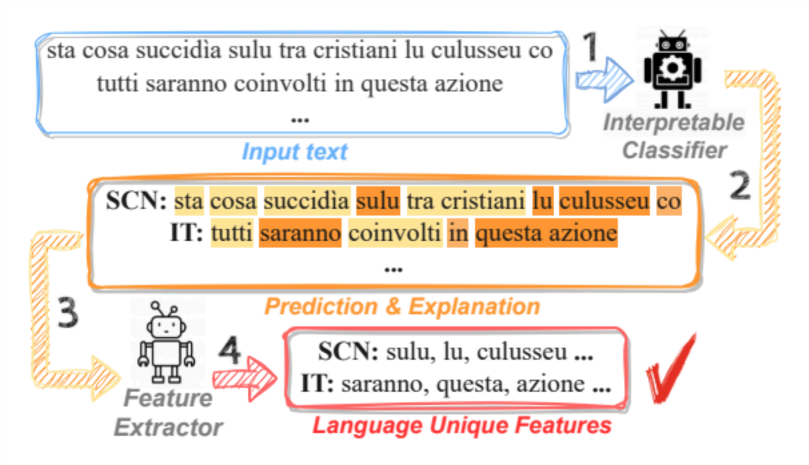


Figure 1: *(1)* given an input text; *(2)* the interpretable dialect classifier return labels (SCN and IT) and explanations; *(3)* the extractor takes the explanations and *(4)* outputs meaningful features to the languages.

Interpretable dialect classifiers



Interpretable dialect classifiers

Core idea – Leave-one-out classification:

- Train a BERT-based dialect classifier using annotated data (e.g. from VarDial shared tasks)
- For each instance of the test set:
    - Record its predicted label and score
    - Remove one word from the instance, record predicted label and score
    - Identify the words that led to sharpest decrease of prediction score (**impact score**)
- Aggregate selected words (**explanations**) over the whole test set

Example (fictive scores)

Ich umarme und bussl Männer ab. ‘I hug and kiss men.’	AT	0.59	
umarme und bussl Männer ab.	AT	0.52	-0.07
Ich und bussl Männer ab.	AT	0.51	-0.08
Ich umarme bussl Männer ab.	AT	0.60	+0.01
Ich umarme und Männer ab.	AT	0.24	-0.35
Ich umarme und bussl ab.	AT	0.49	-0.10
Ich umarme und bussl Männer	AT	0.52	-0.07

Experiments

Xie et al. (2024):

- Mainland Mandarin vs Taiwan Mandarin (FRMT)
- Sicilian vs Italian (ITDI)
- Dutch Low Saxon vs German Low Saxon (LSDC)
- OFL vs all other Low Saxon dialects (LSDC)

Our replication studies:

- Jodel dataset (Hovy & Purschke 2018)
    - Social media data, all kinds of noise
    - 5 classes: AT, CH, Southwest-DE, Southeast-DE, North-DE
- Swiss portion of the Jodel dataset
    - 11 classes: 10 major dialect areas + French/Italian
- Modern Greek dialects (GRDD, Chatzikyriakidis et al. 2023)
    - 4 classes: Northern, Pontic, Cretan, Cypriot

Experiments

Changes to the experimental setup:

- Extension from binary to multi-class settings
- Base model: comparison between XLM-R and language-specific BERTs
- Remove all instances of the same word at once (instead of one at a time and taking maximum impact score)
- Use average impact score to rank the words (instead of additional TF-IDF ranking)

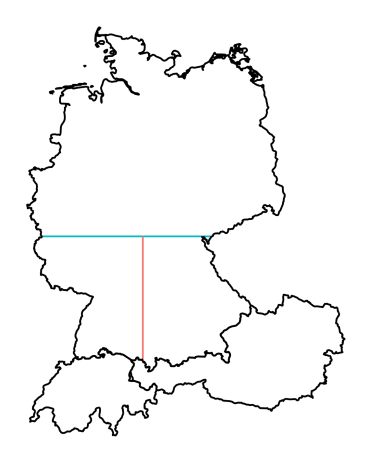
Further potential improvements:

- Tokenization and truecasing
- Word removal vs word masking

Entire Jodel corpus

- 5 classes
- 200k posts per class for training
- 20k posts per class for dev/test
- On average, a post contains 11.5 tokens

Base model	Accuracy
Random	20%
XLM-RoBERTa	47.74%
dbmdz German BERT	47.64%



Expectations

Purschke & Hovy (2019):

Switzerland	esch, ond, vell, gaht, wüki, nöd, besch, emmer, nor, au nöd, verstahn, muen, wükli, dänn, vode, hett, chan, rechtig, staht, sösch, abig, mached, isch de, lüüt, nanig
Northern Germany	ja gut, erstmal, sieht, drauf, vielleicht, mehr, gut, sehen, schonmal, ahnung, bisschen, gesagt, kommt, allerdings, gucken mal, reicht, achja, bestimmt, garnicht, musst, ansonsten, scheinbar, darauf, schon gut, wahrscheinlich
Southern Germany & Austria	afoch, voi, nd, i a, oda, möppes, nimma, is a, mei, gscheid, is, ffm, @vj, hnx, vj, lörres, @vvj, bissl, dummwiekarlsruhe, gibt, vermutlich, lässt, gerade, feuerbach, wobei

- Unsupervised approach, placenames removed

Results

Top 15 explanations per class:

Austria	Switzerland	North-DE	Southwest-DE	Southeast-DE
Klagenfurt	Bern	Bielefeld	Trier	Augsburger
Kärnten	Kei	Mörres	Darmstadt	Erlangen
#schnabeltiereandiemacht	pourquoi	Marburg	Saarland	Leierkasten
wien	grosse	Guckt	@vj	Augsburg
eich	Schweizer	Halle	Tübingen	Passau
Graz	füür	*gucke	Mainz	erlangen
fortgehen	eber	Aachen	KA	Regensburg
Vl	Nid	gucke	Möppes	Bayern
Oasch	contre	Hannover	Stuttgart	Nürnberg
Innsbruck	Basel?	Kiosk	Heidelberg	regensburg
Grazer	uss	guckt	Karlsruhe	Ulm
na,	bim	Köln	Ravensburg	#wudel
#jodelconfession	Bisch	Jena	möppes	SAP
heuer	Huere	Siegen	Vaihingen	Bamberg
Jus	tüür?	guck	Möppes:	#3st

Results

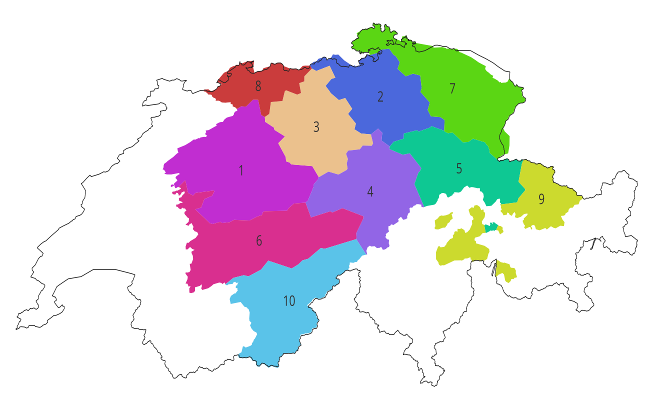
- Place names are prominent in all classes, but mostly for DE classes
- CH: Mostly Swiss German words, also French words
  - Jodel data was collected from all major Swiss cities
  - No language filtering applied
- AT: Dialectal pronunciations (*eich, Oasch*), regionalisms (*fortgehen, heuer*)
- *gucken* as a specifically Northern German lemma
- Jodel-specific “terminology” (*Möppes, Lörres, Mörres*) seems to have originated in Southwest Germany

D. Roemling, A. Miletić & Y. Scherrer: Explainability of Machine Learning Approaches in Forensic Linguistics – A Case Study in Geolinguistic Authorship Profiling. To appear in *Proceedings of NLPAICS*.

A closer look at the Swiss data

We assign the Jodels to one of the 10 major dialect areas according to Scherrer & Stoeckle (2016).

Jodels from outside the German-speaking area (French and Italian) are assigned label 0.



Swiss Jodel subcorpus

Posts per class:

Class	Train	Dev	Test
0	12458	1067	1067
1	12458	1067	1067
2	12458	1067	1067
3	12458	1067	1067
4	2283	513	515
5	10424	1067	1067
6	758	168	173
7	10402	1067	1067
8	12458	1067	1067
9	57	15	21
10	330	82	66

Classification accuracies:

Base model	Accuracy
Random	~9.09%
XLM-RoBERTa	55.68%
dbmdz German BERT	56.80%

Expectations

Purschke & Hovy (2019):

French (0)	t'as, je vais, autant, pour le, que ça, peut être, j'ai, en fait, je pense, …
Bern (1)	geit, viu, gloub, auso, aues, ig, när, ds isch, itz, aube, aui, geng, iz, vilech, ke, ds, nidmau, schnäu, froue, u, ig ha, u när, würklech, angeri, verzeu
Zurich (2+7)	gaht, wüki, nöd, nödmal, vo de, au nöd, verstahn, chan, muen, wükli, gahsch, dänn, vode, hett, isch au, demit, chönd, staht, mached, eifach, abig, isch de, isch scho, git, lüüt
Aarau-Luzern (3)	esch, ond, vell, besch, ech, nor, emmer, au ned, dech, wörkli, wechtig, mech, rechtig, norno, zuekonft, beni, gfonde, brengt, sösch, wössed, drom, esh, dorom, fende, ergendwie
Chur (5)	miar, diar, dia, leba, werda, aswia, wia, aswo, iar, fraua, akli, liabsta, passiart, könna, niamert, muassi, ihar, kriaga, froga, nögsta, muass, vergessa, eba, glauba, guati
Basel (8)	goht, sehni, drnoch, griegsch, syy, keini, usseht, sunsch, miehsam, mol, iebig, öbbis, miesst, au nid, joor, drugge, kha, unseri, friener, isch e, kei, sälbr, joohr, priefig, bitz

Results

No results for classes 6, 9, 10. Few results for class 4:

0	1	2	3	4	5	7	8
femme	Bern	Züri	zom	gu	zglicha	frauefeld	nit
Telegram	biud	nöd!	mech	iher	kanni	mached	Ka
Tu	säuber	gahts	ergendwie	üch	ersta	Freundin	Basel?
Oui	haut	überleit.	get	zämä	posta	wa	basel
!?	aues	züri	dech		luagt	erwischt	ka
quel	Ig	nôd	ehr		gwunna	fühlt	jz
Une	geit	dänn	besch		gera	meisten	dört
Quel	augemein	nachem	geds		Danka	Seit	Sone
Vous	gäud	demit	höt		sacha	Jemand	anderi
J'aime	aube	wegem	ned?		Wuchanend	giz	Gniess

- Few place names
- Generally reasonable explanations
- Class 7 contains significant amounts of Standard German

Conclusions

The proposed approach provides results similar to Purschke & Hovy (2019). But what are the differences?

- Our approach essentially performs dialect identification (DID).
  - Attractive if annotated training datasets exist.
- Our approach yields a trained model that can be applied to new data.
  - Attractive if there is a genuine need for DID.
  - Potential use case: **forensic linguistics**.

Thanks for your attention!

Extracting dialect-specific features from dialect classifiers

Yves Scherrer and Dana Roemling

(Joint work with Aleksandra Miletić and Noëmi Aepli)

ICLAVE 12, Vienna
Panel *Embracing Variability in Natural Language Processing*
10 July 2024