URIEL und Lang2vec

31.01.2024
Huangyan Shan
Profilierungsmodul Computerlinguistik I
Dozent: Dr. Robert Zangenfeind

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übersicht

- Einleitung: URIEL Datenbank und Lang2vec Bibliothek
- Datenquellen
- Vektorentypen
- Distanzen von Sprachen
- Anwendungsmöglichkeiten
- Fazit

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Einleitung

Was ist URIEL und Lang2vec?

- Eine sprachlich typologische Datenbank und ein Hilfswerkzeug für Abfragen
- Binäre Vektoren von typologischen, phylogenetischen und geografischen Merkmalen der Sprachen
- Einheitliche Formaten, Namensgebung, Sprachcodes und Semantik der Merkmale
- Extrahierte Merkmale aus fünf linguistischen Datenbanken
- Vorhergesagte Werte der Merkmale auf der Grundlage geografischer und phylogenetischer Distanzen
- Sechs vorberechnete sprachliche Distanzen

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Datenquellen

- **WALS**: Eine Datenbank von strukturellen phonologischen, grammatischen, lexikalischen Merkmalen der Sprachen
  - Syntax, Phonologie, Geographie
- **SSWL**: Eine Datenbank von morphologischen, syntaktischen und semantischen Merkmalen der Sprachen
  - Syntax, Geographie
- **PHOIBLE**: Eine Sammlung und Normalisierung von sieben phonologischen Datenbanken
  - Phonetische Inventory
- **Ethnologue**: Ein umfassendes Nachschlagewerk, das alle heute bekannten lebenden Sprachen der Welt katalogisiert
  - Syntax (extrahiert durch Textmining von Prosa-Beschreibungen der typologischen Merkmale)
- **Glottolog**: Ein umfassender Katalog der Sprachen, Sprachfamilien und Dialekte der Welt
  - Phylogenie, Geographie

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vektorentypen - Typologische Vektoren

- 103 syntaktische Merkmale
- 28 phonologische Merkmale
- 158 phonetische Inventory Merkmale

Bei fehlenden Werten:

➔ KNN Vorhersage auf der Grundlage geographischer und genetischer Merkmale

➔ Genauigkeit von 92,93 % erzielt

Vector type	#Languages	#Features	#Data points	%Coverage
Syntax (from sources)				
syntax_wals	1808	98	78732	44%
syntax_sswl	230	33	6404	84%
syntax_ethnologue	1336	30	18105	45%
Syntax (averaged over sources)				
syntax_avg	2654	103	94227	34%
Syntax (predicted)				
syntax_knn	7970	103	820910	100%
Phonology (from sources)				
phonology_wals	832	27	14358	64%
phonology_ethnologue	543	8	1017	23%
Phonology (averaged over sources)				
phonology_avg	1296	28	15303	42%
Phonology (predicted)				
phonology_knn	7970	28	223160	100%
Inventory (from sources)				
inventory_phoible_aa	202	158	31916	100%
inventory_phoible_gm	428	158	67624	100%
inventory_phoible_ph	404	158	63832	100%
inventory_phoible_ra	100	158	15800	100%
inventory_phoible_saphon	334	158	52772	100%
inventory_phoible_spa	219	158	34602	100%
inventory_phoible_upsid	334	158	75050	100%
Inventory (averaged over sources)				
inventory_avg	1715	158	270970	100%
Inventory (predicted)				
inventory_knn	7970	158	1259260	100%

Table 3: Typological vectors available in lang2vec, along with the number of languages and features, the number of individual data points, and the percentage of those language/feature pairs for which that data point exists.

(Littell et al., 2017)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vektorentypen – Geographische und phylogenetische Vektoren

Geographische Vektoren:

➔ Der Abstand zwischen Sprachstandorten* und Punkten auf der Erdoberfläche

Phylogenetische Vektoren:

➔ Jede Dimension steht für eine Sprachfamilie oder einen Zweig davon

➔ aus dem Stammbaum der Weltsprachen in Glottolog

	Indo-European	Germanic	West Germanic	Romance	North Germanic
deu	1	1	1	0	0
eng	1	1	1	0	0
fra	1	0	0	1	0
swe	1	1	0	0	1
mlg	0	0	0	0	0

Table 4: Truncated `lang2vec` phylogeny vectors for German, English, French, Swedish, and Malagasy, where 1 represents membership in a particular language family or branch.

(Littell et al., 2017)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Distanzen zwischen Sprachen

Sechs Typen von Distanzen:

- Geographic
- Genetic
- Phonetic Inventory
- Syntactic
- Phonological
- Featural (Kombination)

Language Distances

```
>>> import lang2vec.lang2vec as l2v
>>> l2v.distance('syntactic', 'deu', 'eng', 'chn')
array([[0.  , 0.42, 0.58],
       [0.42, 0.  , 0.57],
       [0.58, 0.57, 0.  ]])
>>> l2v.distance('phonological', 'deu', 'eng', 'chn')
array([[0.    , 0.3277, 0.5456],
       [0.3277, 0.    , 0.5687],
       [0.5456, 0.5687, 0.    ]])
>>> l2v.distance('genetic', 'deu', 'eng', 'chn')
array([[0.    , 0.4286, 1.    ],
       [0.4286, 0.    , 1.    ],
       [1.    , 1.    , 0.    ]])
>>> l2v.distance('geographic', 'deu', 'eng', 'chn')
array([[0. , 0.1, 0.4],
       [0.1, 0. , 0.3],
       [0.4, 0.3, 0. ]])
```

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Anwendung

Auswirkung der typologischen Merkmale für Dependency Parsing Task:

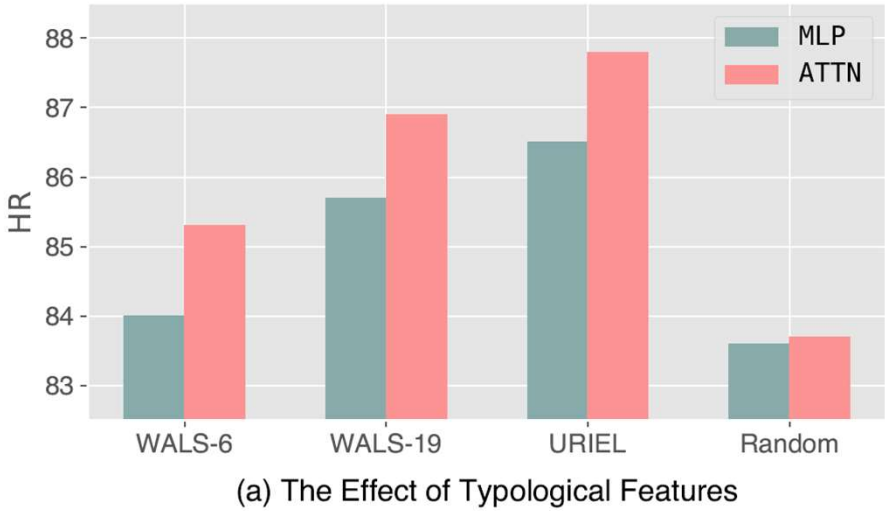
Experiment:

- WALS-6: 6 syntaktische Merkmale aus WALS
- WALS-19: 19 syntaktische Merkmale aus WALS
- URIEL: 103 syntaktische Merkmale aus URIEL
- Random: zufällige Werte

Ergebnis:

- Höheste Verbesserung von LAS der 13 high-resource Sprachen durch URIEL

Typology Guided Multilingual Position Representations: Case on Dependency Parsing



(a) The Effect of Typological Features

(Ji et al., 2023)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Anwendung

Feststellung von Sprachen für Cross-lingual zero-shot transfer Learning:

Experiment:

- Model: mBERT und XLM-R
- Zielsprachen: AR, ZH, FI, HE, HI, IT, JA, KO, RU, SV, TR, EU (aus 8 Sprachfamilien)

Ergebnis:

- Zero-shot Performance für syntaktische Tasks (DEP und POS) hat hohe positive Korrelation mit syntaktischer Ähnlichkeit der Sprachen.

From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers

		SYN		PHON		INV		FAM		GEO		SIZE	
Task	**Model**	P	S	P	S	P	S	P	S	P	S	P	S
DEP	XLM-R	0.77	0.78	**0.83**	**0.77**	0.46	-0.04	0.68	0.61	0.80	0.81	0.62	0.47
	mBERT	**0.92**	**0.91**	0.79	0.74	0.55	-0.01	0.76	0.62	0.64	0.69	0.79	0.59
POS	XLM-R	0.68	0.79	**0.81**	**0.81**	0.38	0.02	0.58	0.74	0.80	0.73	0.54	0.46
	mBERT	**0.90**	**0.87**	0.86	0.81	0.57	0.02	0.82	0.80	0.66	0.72	0.47	0.39
NER	XLM-R	0.49	0.49	**0.80**	**0.83**	0.27	0.14	0.47	0.55	0.77	0.81	0.37	0.35
	mBERT	0.60	0.74	**0.81**	**0.84**	0.34	-0.04	0.53	0.58	0.59	0.73	0.42	0.38
XNLI	XLM-R	**0.88**	**0.90**	0.29	0.27	0.31	-0.11	0.63	0.54	0.54	0.74	0.70	0.76
	mBERT	**0.87**	0.86	0.21	0.08	0.29	0.04	0.61	0.47	0.55	0.67	0.77	**0.91**
XQuAD	XLM-R	0.69	0.53	**0.85**	**0.81**	0.62	-0.01	**0.81**	0.54	0.43	0.50	**0.81**	0.55
	mBERT	0.84	0.89	0.56	0.48	0.55	0.22	0.79	0.64	0.51	0.55	**0.89**	**0.96**

Table 2: Correlations between zero-shot transfer performance with mBERT and XLM-R for different downstream tasks with linguistic proximity features (SYN, PHON, INV, FAM and GEO) and pretraining size of target-language corpora (SIZE). Results reported in terms of Pearson (P) and Spearman (S) correlation coefficients.

(Lauscher et al., 2020)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Fazit

Vorteile:

- Umfassende typologische Merkmale
- Einheitliche Formaten und Namensgebung
- Einfach nachzuschlagen (mithilfe Lang2vec Bibliothek)
- Verfügbare vorberechnete Distanzen zwischen Sprachen

Nachteile:

- Unvollständige Einträge (weniger als 50% vorhanden)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Literatur

- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499, Online. Association for Computational Linguistics.
- Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning Language Representations for Typology Prediction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.
- Chris Collins and Richard Kayne. 2011. Syntactic Structures of the World's Languages. New York University, New York.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. Glottolog 2.6. Max Planck Institute for the Science of Human History, Jena.
- Matthew S. Dryer and Martin Haspelmath. 2013. The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology.
- M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. Ethnologue: Languages of the World, Eighteenth edition. SIL International, Dallas, Texas.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
- Steven Moran, Daniel McCloy, and Richard Wright. 2014. PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Tao Ji, Yuanbin Wu, and Xiaoling Wang. 2023. Typology Guided Multilingual Position Representations: Case on Dependency Parsing. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13524–13541, Toronto, Canada. Association for Computational Linguistics.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vielen Dank für Ihre Aufmerksamkeit