Probing von großen Sprachmodellen auf morphologische und morphosyntaktische Merkmale

Referat im Profilierungsmodul Computerlinguistik I - Dr. Robert Zangenfeind
Wintersemester 2023/24 @CIS, LMU München
Xaver Maria Krückl, 17.01.2024

Gliederung

- Einführung
- Morphologisches & Morphosyntaktisches Probing
  - Probing (mit multiplen Labels)
  - Monolinguale Experimente
  - Multilinguale Experimente
  - Cross-Linguale Experimente
- Fazit
- Bibliographie
- Appendix

Einführung

Morphologie in (großen) neuronalen Sprachmodellen

- genereller Konsens in der Forschung:
  neuronale Sprachmodelle **erfassen morphologische** (morphosyntaktische) **Merkmale** und Abhängigkeiten
- auch ohne entsprechende Anleitung beim Training!
- aber: **morphologisch komplexe/reichhaltige Sprachen** stellen die Modelle vor **Herausforderungen**:
  - komplexe Kongruenz Muster (agreement patterns)
  - limitiertes Vokabular vs. große Vielfalt flektierter Formen

(Shapiro et. al. 2021, Matthews et. al. 2018, Vania et. al. 2018, Park et. al. 2021)

Einführung

Überblick über den Forschungsstand #1 (Matthews et. al. 2018)

- erste Analysen bereits anhand RNN basierter Modelle
  - Probleme: nicht alle Wortformen können aus feststehendem Vokabular erzeugt/repräsentiert werden
  - Lösung? **buchstaben-basierte** Modelle (character based) aber: linguistisch sehr **naiv** (muss lernen was Wörter sind)
  - beser: Kombination verschiedener generativer Modellarchitekturen: buchstaben-, wort- und morphembasiert, letzteres trainiert mit Zugriff auf handgeschriebenem morphologischen Analysator

Einführung

Überblick über den Forschungsstand #2 (Vania et. al. 2018)

- weitere Analysen zu buchstaben-basierten Modellen:
  - “erlernen” Morphologie komplexer Sprachen - können Wortformen außerhalb des Trainings-Vokabulars repräsentieren (keine Probleme mit Orthographie)
  - aber: Probleme bei **Kasussynkretismus** (Formgleichheit, zeigt sehr naives “Erlernen” von Wörtern)
  - ebenfalls deutliche Verbesserung durch Hinzuziehen explizit morphologischen Wissens beim Training

Einführung

Überblick über den Forschungsstand #3 (Park et. al. 2021)

- Analyse des Einflusses von Morphologie in **multilingualen** Sprachmodellen (LSTM basiert)
  - Evaluierung ähnlich zum Probing hier (*surprisal*)
  - Einfluss des **Encodings** auf morphologisches Wissen?
  - ➢ Byte-Pair-Encoding hemmt das Erlernen morphologischer Merkmale (BPE zerlegt Token nach Teilhäufigkeiten)
  - Verbesserung durch linguistisch motivierte Kodierung und Segmentierung (endliche Automaten/*Morfessor* Tool)

Einführung - Motivation

- Morphologische Überwachung (supervision) kann (mehrsprachige) Sprachmodelle diesbezüglich verbessern
- **aber**: auch Modelle die ohne morphologische Induktion trainiert wurden erfassen morphosyntaktische Phänomene!
- ✦ Wie gehen große, modernere neuronale Sprachmodelle mit der Komplexität morphologisch reichhaltiger Sprachen um?
  - ➔ Erforschung der Repräsentation von morphologischen und morphosyntaktischen Merkmalen in solchen Modellen
  - ➔ ohne Überwachung/Wissensbasis beim Training
- wichtig für Bereiche wie maschinelle Übersetzung, Question-Answering und Sprachgenerierung

(Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels (Shapiro et. al. 2021)

1. Einführung eines **effizienten “Probing” Paradigmas**
   - ➔ zur Analyse mehrerer morphosyntaktischer Merkmale in (mehrsprachigen) neuronalen Sprachmodellen
   - ➔ gezeigt anhand von **multilingual-BERT** (mBERT) und 7 typologisch diversen Sprachen
2. **Evaluierung der Probes** auf 6 ebenso diversen Sprachen
3. Veröffentlichung des Codes und der Vorhersagen
   - ➔ Grundlage und ein Leitfaden für weitere Probing-Ansätze und tiefergehende morphologische Merkmals-Analysen

Probing

- Technik zum **Erfassen linguistischer Merkmale** in **neuronalen Sprachmodellen**
- Trainieren eines **Klassifikators auf** einem **vor-trainierten** neuronalen Sprachmodell (den Embeddings des Modells)
- Training/Evaluation des Klassifikators setzt nach linguistischen Merkmalen gelabelten Trainings- und Testdaten voraus!
  - hier: morphologische bzw. morphosyntaktische Merkmale
  - konkret: eine Menge von Sätzen in denen jedes Wort entsprechend dieser Merkmale die es trägt gelabelt ist
- Verwendung des trainierten Klassifikators zur Vorhersage der morphosyntaktischen Eigenschaften von ungesehenen Sätzen

(Shapiro et. al. 2021)

Probing

- Problem: Erinnerungseffekte (memorization/overfitting) von sog. “extraktiven” Probes - zu stark an Trainingsdaten angepasst
- wollen Probes die “selbst lernen”, also diese Effekte minimieren
- erreichbar durch:
  - Begrenzen der Probe-Komplexität (durch höheren Dropout)
  - Reduzieren der Menge an Trainingsdaten
  - Verwenden einfachere Modelle Architekturen des Klassifikators (z.B. nur eine Lineare Layer anstatt eines Multi-Layer Perzeptrons)

(Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels

- **“spezielles” Probing Paradigma** hier:
  - alle **morphosyntaktische Merkmale** zusammengefasst in nur **einer “diagnostischen” Task/Probe**
  - ergänzende **“kontroll” Task** zum **leiten und interpretieren der Probes** (Tests)
    - ➢ trainiert, um **randomisierte Outputs** aus den selben Embeddings wie die diagnostische Task vorherzusagen
- ***Selektivität***: Maß das die Differenz der Performanz zwischen diagnostischer- and kontroll-Task angibt
- **Interpretation**: je größer die Selektivität, desto mehr morphosyntaktische Information ist in der Eingabe der Probe (den vor-trainierten Embeddings) enkodiert

(Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels

- benötigt **morphosyntaktisches Tagging** für **multiple Labels!**
  - ➔ Verwendung morphologisch annotierter Daten des **Universal Dependencies** (UD) (z.B. de Marneffe et. al. 2021)
- morphosyntaktisches Tagging als eine Task auf **Wort-Ebene**:
  - ➔ ein Token kann **multiple Merkmals-Label** erhalten: (z.B. *person=1*, *number=sing*)
  - ➔ Token werden **multi-hot enkodiert**
  - ➔ erlaubt auch das enkodieren von mehrteiligen (multi-valued) Merkmalen (z.B. *gender=fem,masc*)
  - ➔ erlaubt eine **genauere Analyse** der erlernten **Dependenz-** und **Merkmals-Kookkurrenz Muster**, keine Isolation der Merkmale (Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels

Multi-Hot Encoding - Beispiel:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	SCONJ	VERB	Case=Acc	Case=Gen	PronType=Art	PronType=Prs	Person=1	Person=2	Person=3	Gender=Fem	Gender=Masc	Number=Dual	**Number=Sing**	**Number=Plur**	Tense=Fut	Tense=Past	Mood=Imp	VerbForm=Inf	VerbForm=Part	
”she” -	[0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	**1**	**1**	0	0	**1**	0	0	0	0	0	0]	...
”they” -	[0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	0	0	**1**	**1**	0	0	0	**1**	0	0	0	0	0]	...

(Shapiro et. al. 2021: 4487)

- zwei “gold” Vektoren - für das Personalpronomen in der dritten Person - **Singular/Plural**- Femininum (unvollständig)
- wie bei gewöhnlichem One-Hot Encoding: **1** wenn das Token (Pronomen) das entsprechende Label aufweist, **0** wenn nicht

Morphosyntaktisches Probing mit multiplen Labels

- F = Menge an **Merkmals-Labels** $\{f_1, ..., f_{|F|}\}$ (morphosyntaktische Eigenschaften)
- V = **Vokabular** (Wort **Types**)
- $s = s_1 ... s_{|s|}$ = ein spezifischer **Satz**
- r_i = die **kontextualisierte Repräsentation** jedes **Token** s_i in mBERT, sodass $s_i \in V$
- **Input** für eine **Probe**: das Embedding $r_i \in R^d$
- wie für multilabel morphosyntaktisches Tagging: **Ziel-Output** jedes **Embeddings** r^i = **multi-hot** enkodierter **Vector** y^i
- $y^i = y^i_1 ... y^i_{|F|}$ (Vektor über alle Merkmals-Labels)

(Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels

- Label-Sets:
  - insgesamt **166** unterschiedliche **Merkmals-Label** (s. Appendix)
  - **unterschiedliche** Label-Sets für die Probe jeder Sprache in **monolingualen** Szenarios
  - ein **teilweise aggregiertes** Label-Set für **multilinguale** Probe
- Evaluation der Merkmale:
  - **individuell**: Berechnung von Precision, Recall und F1 für jedes Merkmals Label
  - **holistisch**: Berechnung des micro-average F1 einer Probe mittels der TPs, FPs und FNs über alle Merkmale

(Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels

- Modell:
  - "**BERT-Base, Multilingual Cased**" Modell aus der Huggingface Transformers Bibliothek (mBERT)
  - 110M Parameter, 12 Transformer Layers, jede mit 12 Attention Heads, Hidden Size 768
  - trainiert auf Wikipedia Daten der top 104 Sprachen, cross-linguales Vokabular von 100K Wörtern
  - Fokus beim Training der Probe-Klassifikatoren auf **geradzahlige Layers** (z.B. mBERT-0, mBERT-2, ...)

  ➔ vorherige Studien zeigen **layer-by-layer Trends**: linguistische Repräsentationen in BERT darin am klarsten

(Shapiro et. al. 2021)

Morphosyntaktisches Probing mit multiplen Labels

- Training der linearen Probing Klassifikatoren:
  - **“einfrieren”** des **vor-trainierten mBERT Modells**
  - Extrahieren der Embedding Layer und der entsprechenden Transformer Layers
  - Ausführen des **Taggings** für **multiple Merkmals-Label** auf den Embeddings aller Training/Test-Token aller Layers
  - Sigmoid Aktivierungsfunktion
  - Loss: mean binary cross-entropy
  - Training: 50 Epochen - auf besten Validation-Loss
  - PyTorch **Hyperparameter**: Optimizer Adam, Lernrate 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e\text{-}08$, kein Dropout

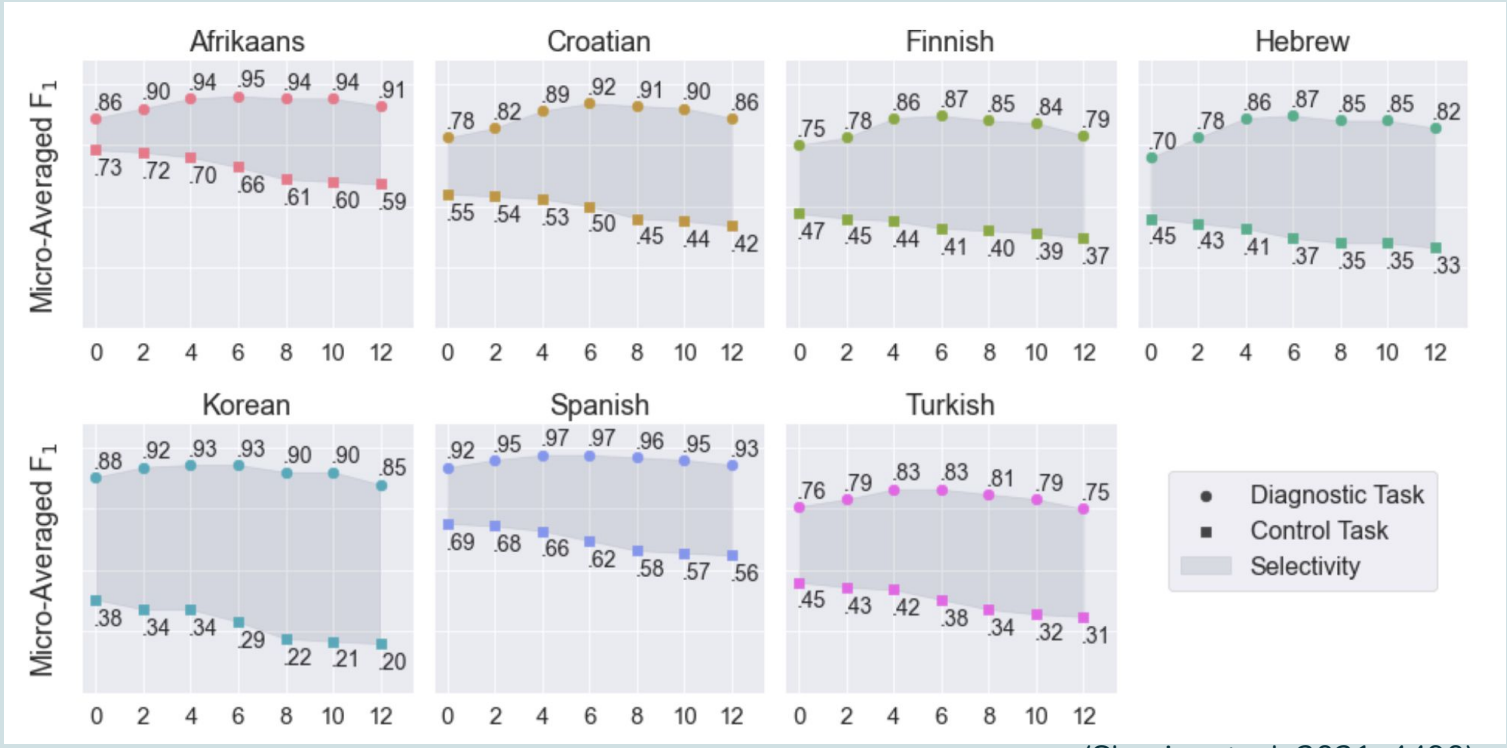
(Shapiro et. al. 2021)

Monolinguale Experimente

- Training und Evaluierung von **individuellen diagnostischen Probes für 7 Sprachen**: *Afrikaans*, *Kroatisch*, *Finnisch*, *Hebräisch*, *Koreanisch*, *Spanisch* und *Türkisch* (800 Sätze pro Sprache)
- micro-average F1 Scores zeigen: mBERT ermöglicht eine **einfache Extraktion** morphosyntaktischer Merkmale
- Probe der **6. mBERT Layer** mit dem besten Ergebnis über alle Sprachen - konsistent mit vorherigen Ergebnissen für das generelle BERT Modell bzgl. ähnlicher linguistischer Tasks
- beste F1 Scores zwischen 0.83 und 0.97, hohe Selektivität
- visualisiert:

(Shapiro et. al. 2021)

Monolinguale Experimente



(Shapiro et. al. 2021: 4490)

Monolinguale Experimente

- morphologisch relevante Information werden also **bis zur 6. Layer** in mBERT enkodiert
- danach scheint Information dazu abzunehmen - weitere Layers beschäftigen sich mit der Speicherung anderer Muster
- *Afrikaans* und *Spanische* Probes schneiden am besten ab, aber *Türkisch* am schlechtesten?
  Vorsicht: unterschiedliche Größe des Trainingssets (Tokens der Sätze) und unterschiedliche Menge an Merkmalen - Sprachen **nicht untereinander vergleichbar!**
- diagnostische Probes übertreffen control Probes deutlich
- Trend: **Verbesserung der Selektivität** in höheren Layers

(Shapiro et. al. 2021)

Monolinguale Experimente

- mBERT erlernt also (eine Art) morphosyntaktischen Wissens
- aber: Scores geben keine direkten Einblicke …
  - welche Aspekte der Morphosyntax repräsentiert sind
  - in das Zusammenspiel zwischen verschiedenen morphosyntaktischen Eigenschaften
  - wie variabel mBERT in der Erfassung einzelner Merkmalswerte ist
- dabei doch Vorteil des Probings mit multiplen Labels: Erkennbarkeit feiner morphosyntaktischer Beobachtungen die mehrere Merkmale betreffen?!

(Shapiro et. al. 2021)

Monolinguale Experimente

Fallstudie: (verdeckte) Hebräische Artikel

- Verwendung der 6. Layer in mBERT (beste F1 Scores)
- Hebräisch: ambige Orthographien und Mehrwort-Tokens (MWTs)
- übliche Struktur eines MWTs z.B.: ADP-(DET)-NOUN
- DET als bestimmter Artikel, kann fehlen (ist verdeckt/implizit)
- Schwierigkeit für mBERT, PronType=Art zu erkennen, wenn DET als bestimmter Artikel verdeckt ist
  - vereinfacht durch Dependenzmuster zwischen ADP und NOUN
  - sowie durch Beispiele mit präsentem DET (nicht verdeckt)

➢ ersichtlich aus **qualitativer! Analyse** der falsch negativen Klassifikationen (FNs)

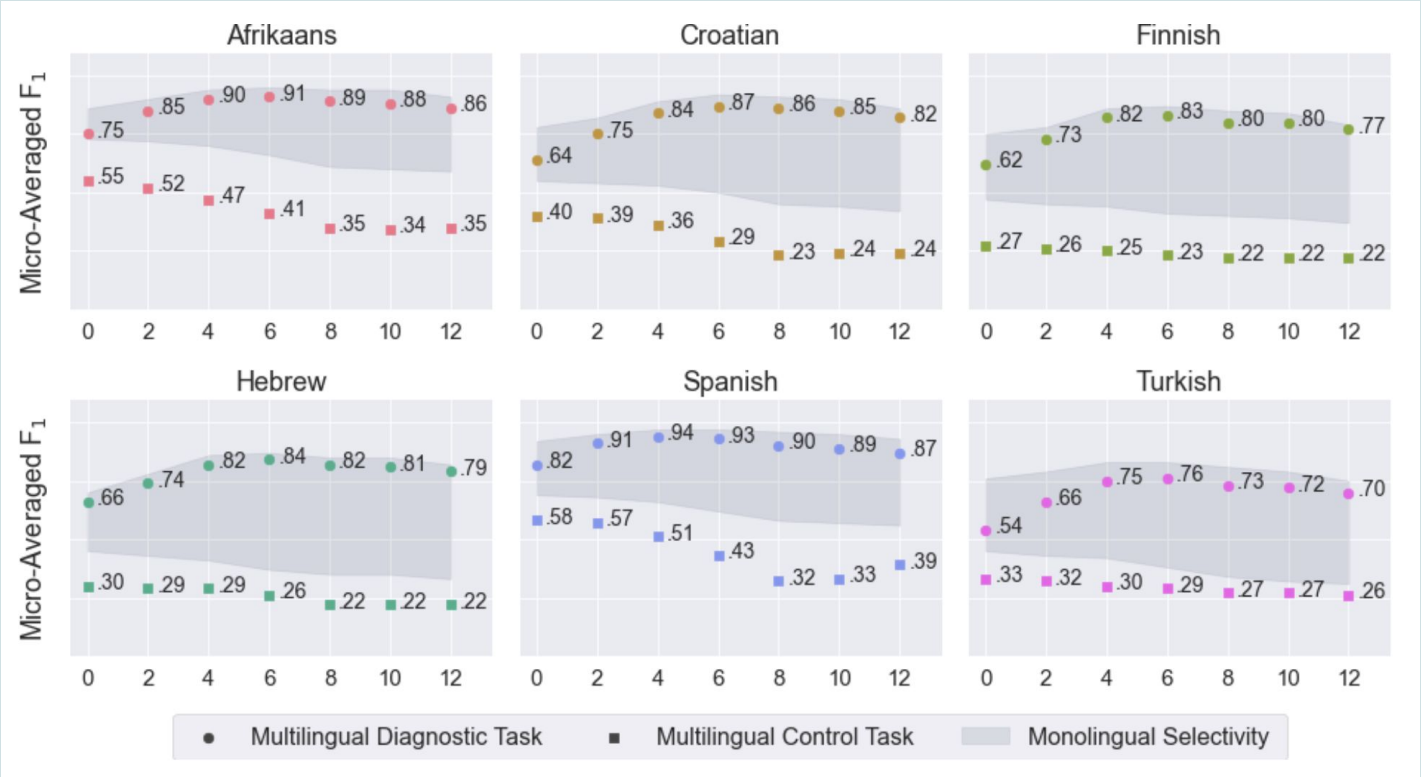
(Shapiro et. al. 2021)

Multilinguale Experimente

- Können die individuellen monolingualen Probes durch **eine einzelne multilinguale Probe ersetzt** werden?
  - ➔ Training einer multilingualen Probe auf einer **gemischten Kombination** von Trainingssets aus *Afrikaans*, *Kroatisch*, *Finnisch*, *Hebräisch*, *Koreanisch*, *Spanisch* und *Türkisch* (4.800 Sätze)
  - ➔ Probe extrahiert eine **aggregierte Teilmenge von Merkmalen** die die monolingualen Probes erfassen würden
- **unabhängiges Beurteilen** für jede Test Sprache
- Ergebnis insgesamt: **leichte Verschlechterung der Performanz**
- aber: **bessere Selektivität** als die monolingualen Pendants

(Shapiro et. al. 2021)

Multilinguale Experimente



(Shapiro et. al. 2021: 4492)

Multilinguale Experimente

- **multilinguale Task komplexer**! die Probe muss:
  - ➔ die Anforderungen mehrerer Sprachen ausbalancieren
  - ➔ die Merkmale aus einer größeren Datenmenge extrahieren
- **marginal schlechtere Performanz** gegenüber monolingualen Pendants - warum?
  - ➔ **Distribution linguistischer Eigenschaften** in den Trainingsdaten unterscheidet sich **cross-lingual**
  - ➔ **gegensätzliche Trainingssignale** an die Probe
  - ➔ weniger (zu wenige) Erinnerungseffekte
- multilinguale Probes dennoch eine **ausdrucksstärkere Anzeige** für linguistische Repräsentationen (höhere Selektivität!)

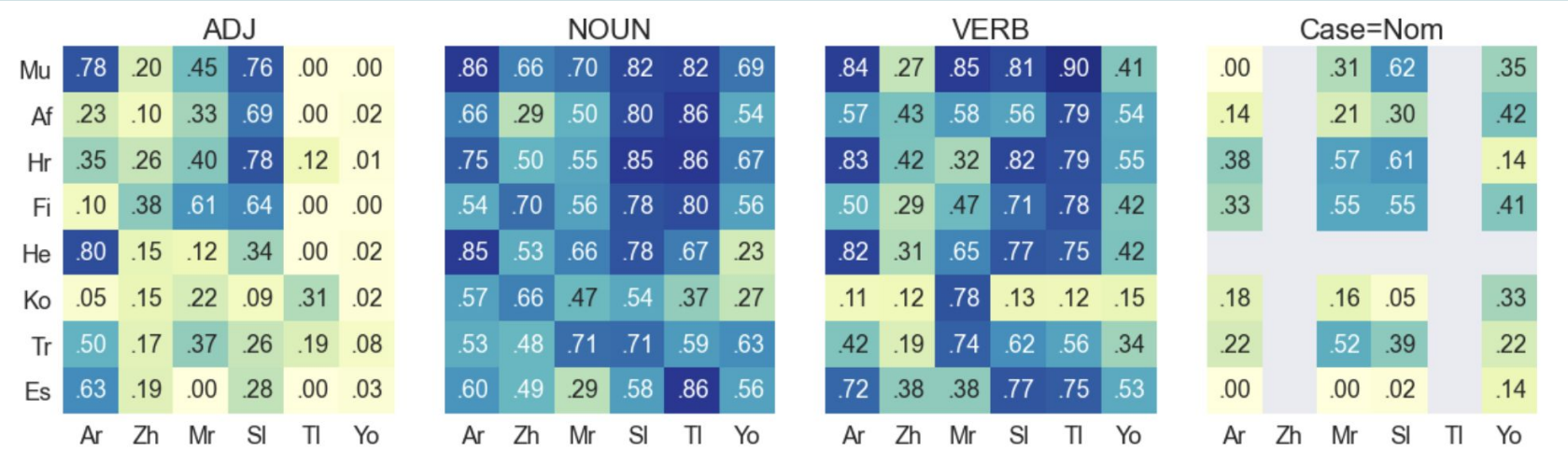
(Shapiro et. al. 2021)

Cross-Linguale Experimente

- Welche morphosyntaktischen Merkmale sind **cross-lingual** (sprachübergreifend) ähnlich enkodiert?
- Evaluation von **monolingualen** und **multilingualen** Probes auf **6** Sprachen, die **nicht Teil** der **Trainingsdaten** waren: *Arabisch*, *Chinesisch*, *Marathi*, *Slovenisch*, *Tagalog* und *Yoruba*
- Fokus nur auf die **6. mBERT** Layer
- Untersuchung einer kleineren **Teilmenge** der Merkmale
- wenn die Probe ein Merkmals-Label erfolgreich extrahiert, hat das Sprachmodell diese linguistische Eigenschaft cross-lingual erkannt
- Ergebnisse:

(Shapiro et. al. 2021)

Cross-Linguale Experimente



- Performanz bzgl. dem Extrahieren von *Nomen* und *Verben* relativ gut
- *Adjektive* sind sprachübergreifend weniger kohäsiv repräsentiert
- kein *Nominativ* als Kasus im *Hebräischen*, *Chinesischen* und in *Tagalog* (nach UD) - keine Ergebnisse hier möglich

(Shapiro et. al. 2021: 4493)

Cross-Linguale Experimente

- mBERT enkodiert die Eigenschaft *Nomen* oder *Verb* zu sein **sprachübergreifend** sehr ähnlich
  ➔ es scheint also eine Art **Konzept** zu *Nomen* und *Verben* zu geben, das über individuelle Sprachen **hinausgeht**
- Eigenschaft *Adjektiv* hingegen nicht in gleichem Maße mit ihren Pendants in anderen Sprachen zusammenhängend
- möglicher Grund für ähnlich enkodierte Merkmale: Kontakt zu multiple Sprachen stellt **reichhaltigere/genauere Verbindungen** in mBERT's Embeddings her

(Shapiro et. al. 2021)

Cross-Linguale Experimente

- **verwandte Sprachen**: **monolinguale Probes** schneiden **besser** auf unbekannten Sprachen ab als die Multilinguale
- *Hebräische* Probes sind den *Arabischen* am ähnlichsten

➔ **Semitische** Sprachen, aber: unterschiedliche **Schrift!**

➔ Repräsentation von mBERT basiert **nicht nur** auf **strukturellen** Ähnlichkeiten sonder auch auf **familiären**

- *Kroatische* mBERT-6 Probe erreicht 0.70 F1 auf *Slovenisch*

➔ sogar 0.95 F1 für Mood=Cnd und 0.86 F1 für Mood=Ind

➔ beide Sprachen teilen mehrere Auxiliare die *Modus* markieren

(Shapiro et. al. 2021)

Fazit

- **morphosyntaktische Repräsentationen** aus multilingualen Wordembeddings (mBERT) mittels **Probing extrahierbar**
- ermöglicht hauptsächlich **holistische** aber auch **feinere Analysen individueller Merkmale** & **Dependenzphänomene** (agreement)
- zeigt dass **Merkmale auch cross-lingual** repräsentiert sind
- Paper ermutigt **unterschiedliche multilinguale Sprachmodellen** mit dem präsentierten Paradigma zu proben
  Aber: ist das so einfach möglich?
  - ➔ Problem an die derzeit größten Modelle zu kommen, v.a. falls diese nicht Open Source sind (z.B. GPT-3)
  - ➔ Anspruch Multilingualität! (z.B. bei Llama 2 nicht gegeben)

(Shapiro et. al. 2021)

Fazit - Kritik

- kann also überhaupt von “großen” Sprachmodellen die Rede sein?
  - ➢ mBERT aus heutiger Sicht nicht mehr wirklich groß
  - ➢ als Teil der BERT Familie basierend auf **masked language modeling** - deutlich andere Ergebnisse für Modelle basierend auf **next word prediction** erwartbar? (Devlin et. al. 2018)
- leider keine Ergebnisse für welche Sprache welche konkreten morphologischen und morphosyntaktischen Phänomene im Modell wie gut repräsentiert sind
  - ➢ nur holistische Angabe der Klassifikationsergebnisse
  - ➢ Durchführung qualitativer Analysen scheint schwieriger zu sein
- keine genauen Angaben zu Architektur und Güte der Klassifikatoren

Bibliographie #1

Jacob **Devlin**, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, 2018.

Daniel **Jurafsky** and James H. Martin. "Speech and Language Processing (Draft of December 29, 2021)." (2021).

Marie-Catherine **de Marneffe**, Christopher Manning, Joakim Nivre, Daniel Zeman (2021): Universal Dependencies. In: Computational Linguistics, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.

Austin **Matthews**, Graham Neubig, and Chris Dyer. 2018. Using Morphological Knowledge in Open-VocabularyNeural Language Models. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1435–1445, New Orleans, Louisiana. Association for Computational Linguistics.

Hyunji Hayley **Park**, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis. Transactions of the Association for Computational Linguistics, 9:261–276.

Naomi **Shapiro**, Amandalynne Paullada and Shane Steinert-Threlkeld. A multilabel approach to morphosyntactic probing. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4486–4524. Association for Computational Linguistics, November 2021.

...

Bibliographie #2

Thomas **Wolf**, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest and Alexander Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

Clara **Vania**, Andreas Grivas, and Adam Lopez. 2018. What do character-level models learn about morphology? The case of dependency parsing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2573–2583, Brussels, Belgium. Association for Computational Linguistics.

Bildquellen:

Slide 1: https://www.midjourneyai.ai/de

Appendix - Teilübersicht POS Labels

Table B1: The monolingual probes extracted different sets of features, while the multilingual probes extracted a semi-aggregated subset of these features (in bold under "Feature Labels").

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
ADJ	✓	✓	✓	✓	✓	✓	✓
ADP	✓	✓	✓	✓		✓	✓
ADV	✓	✓	✓	✓	✓	✓	✓
AUX	✓	✓	✓	✓	✓	✓	✓
CCONJ	✓	✓	✓	✓	✓	✓	✓
DET	✓	✓		✓	✓	✓	✓
NOUN	✓	✓	✓	✓	✓	✓	✓
NUM	✓	✓	✓	✓	✓	✓	✓
PART	✓	✓			✓	✓	
PRON	✓	✓	✓	✓	✓	✓	✓
PROPN	✓	✓	✓	✓	✓	✓	✓
SCONJ	✓	✓	✓	✓		✓	
VERB	✓	✓	✓	✓	✓	✓	✓

fortlaufend...

(Shapiro et. al. 2021: 4504)

Appendix - Teilübersicht Kasus Labels

Case=Abe			✓				
Case=Abl			✓				✓
Case=Acc	✓	✓	✓	✓	✓	✓	✓
Case=Ade			✓				
Case=Advb					✓		
Case=All			✓				
Case=Com			✓			✓	
Case=Comp					✓		
Case=Dat		✓				✓	✓
Case=Ela			✓				
Case=Equ							✓
Case=Ess			✓				
Case=Gen		✓	✓	✓	✓		✓
Case=Ill			✓				
Case=Ine			✓				
Case=Ins		✓	✓				✓
Case=Loc		✓					✓

fortlaufend...

(Shapiro et. al. 2021: 4500)

Appendix - Übersicht Sprach Daten

Table A1: Composition of the training and evaluation data for the monolingual and multilingual probes.

Language	Genus	$\|F\|$	Train		Dev		Test	
			Sentences	Tokens	Sentences	Tokens	Sentences	Tokens
Afrikaans	Germanic	53	800	21,160	194	5,317	425	10,065
Croatian	Slavic	66	800	17,811	960	22,292	1,136	24,260
Finnish	Finnic	89	800	10,786	1,363	18,311	1,553	21,069
Hebrew	Semitic	53	800	16,061	484	8,358	491	8,829
Korean	Korean	35	800	13,177	100	1,679	100	1,728
Spanish	Romance	63	800	24,345	1,654	52,161	1,719	52,429
Turkish	Turkic	64	800	8,244	983	9,768	981	9,794
Multilingual	n/a	72	4,800	98,297	5,638	116,207	n/a	n/a

(Shapiro et. al. 2021: 4499)