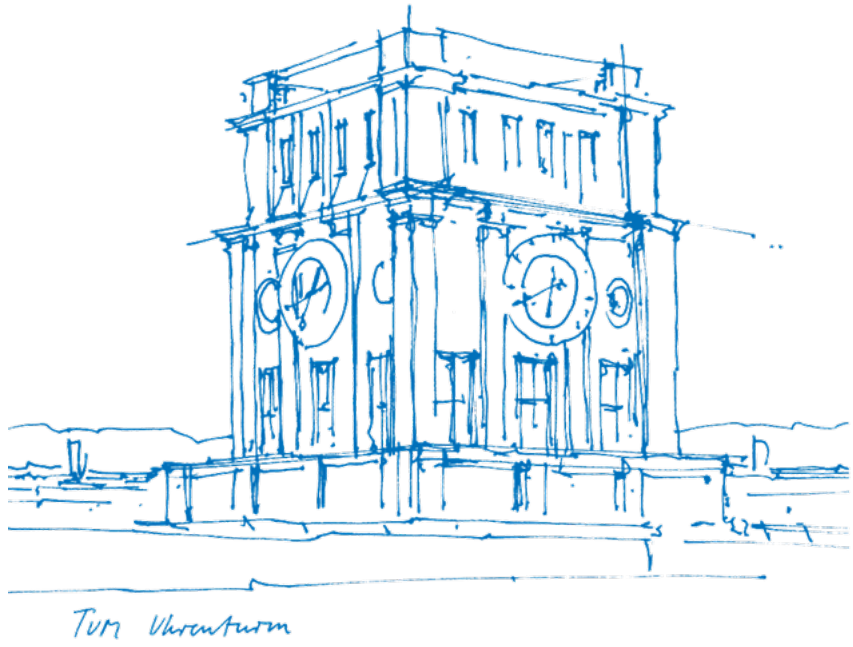
1st Online Workshop on NLP for Language Activists

Alexander Fraser, Shu Okabe
Technische Universität München



Introduction to the Chair for Data Analytics and Statistics

- Research on machine translation, multilingual NLP, multilingual chatbots
  - About 10 researchers (PhD students and postdocs, who, btw, authored most of these slides, thanks! And thanks to Shu particularly!)
  - Previous work with language activists on languages like Upper and Lower Sorbian (hi!), Hiligaynon, Occitan, ...

- Funding from European Research Council on collecting multilingual data (subject of Shu's talk), and new Advanced Grant on "Evaluating and Programming Intelligent Chatbots for Any Language (EPICAL)"

Purpose of this part of the workshop

- Basic introduction to NLP technologies (and a few issues) that language activists will hopefully be interested in
- Not really aimed at NLP researchers, who might find the technical part of this presentation to contain information that they already mostly know
- But here are a few things that are important for NLP researchers:
  - The sustainability of a language is 100% determined by whether children use the language
  - Issues of language sovereignty are important (several slides on this)
  - Government policy is important (out of scope here)
  - Machine translation and other NLP technologies are not important, but “nice to have” (can be useful and also prestige is involved, see first bullet)

Language sovereignty: on data access

- Policy on language technology and data can vary from community to community
  - Most languages lie on the open-access side through massive data from the Internet, such as Common Crawl, Wikipedia, or OPUS (Tiedemann, 2011)
  - Caution when it comes to Indigenous languages
    "*After generations of exploitation, Indigenous people often respond negatively to the idea that their languages are data ready for the taking.*" (Bird, 2020)

Example of the Kaitiakitanga License for Māori (`mri`) in New Zealand.

Direct quotes from the Kaitiakitanga section (*guardianship*) of the website explaining its purpose:

- "Te Hiku Media have developed a Kaitiakitanga licence, which states that data is **not owned** but is **cared for** under the principle of kaitiakitanga and any benefit derived from data flows to the source of the data."
- "Māori data will not be openly released, but requests for access to the data, or for the use of the tools developed under the platform, will be managed using tikanga Māori."
- "Research on other indigenous languages that is carried out under this platform will be for the primary benefit of those peoples."
- "Machine learning software, that is independent of the language communities, will be made openly accessible where appropriate."

Language sovereignty: on NLP collaboration

- NLP field is becoming more aware of the ethical factors in a multilingual context, especially for Indigenous languages: (Bird, 2020), (Schwartz, 2022), (Liu et al., 2022), (Mager et al., 2023)

Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers

Survey of 22 language stakeholders on the development of Machine Translation systems for their language

- Language sharing: most support for unrestricted access, but some prefer to have some limitations
- Impact of MT: Overall support for developing MT systems, but fears of some risks (e.g., language standardisation)
- Collaboration: Importance of involving community members during development
- Data access and ownership: While most agree on public access, views on data ownership depend on the community

- Language sovereignty should be considered case-by-case.
  - Despite overall positive impressions, precaution is needed.
- Willingness to foster collaboration between language communities and NLP practitioners.

Language sovereignty: An Overview

Definition

Language sovereignty refers to the right and ability of a language community to maintain, protect, and promote the use of their native language. It encompasses the political, cultural, and social aspects of language preservation and development.

- Importance
  - **Cultural Preservation:** Language is a key element of cultural identity. Preserving a language means preserving unique worldviews, traditions, and histories.
  - **Political Power:** Language sovereignty is often tied to political independence, as it allows communities to maintain control over their own narratives and institutions.
  - **Social Justice:** Revitalizing endangered languages can address historical injustices, such as colonialism and cultural erasure.
- Case Studies
  - **Zapotec Language in Mexico:** Efforts to reclaim and revitalize the language[1].
  - **Zapara Language in Ecuador:** How language revitalization has supported self-determination[2].

[1] Blog Post: https://www.momentslog.com/culture/mexican-indigenous-languages-revival-preservation-efforts-and-language-revitalization-programs

[2] Paper Reading: https://www.jstor.org/stable/30131233

Language sovereignty: Data and Resources

- **UNESCO Atlas of the World's Languages in Danger**
  - Comprehensive resource on endangered languages
  - UNESCO Atlas[3]
  - *Paper:* Moseley, Christopher (ed.). 2010. *Atlas of the World's Languages in Danger, 3rd edn.* Paris, UNESCO Publishing.
- **Ethnologue: Languages of the World**
  - Extensive database on languages worldwide
  - Ethnologue[4]
  - *Paper:* Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2023. *Ethnologue: Languages of the World, 27th edn.* Dallas, SIL International.
- **Endangered Languages Project**
  - Collaborative project documenting endangered languages
  - Endangered Languages Project[5]
- **SIL International Language Documentation**
  - Resources for language documentation and preservation
  - SIL International[6]

[3] http://webarchive.unesco.org/20170505185847/http:/www.unesco.org/languages-atlas/index.php

[4] https://www.ethnologue.com/index.html

[5] https://www.endangeredlanguages.com

[6] https://www.sil.org/resources/publications/ethnologue

Chair for Data Analytics & Statistics
School of Computation, Information and Technology (CIT)
Technical University of Munich

Unicode

Chair for Data Analytics & Statistics
TU München

Unicode is important for Low-Resource NLP

- **Unicode** is a universal character encoding standard that provides a unique number for every character, no matter the platform, program, or language.
- It ensures the consistent representation of text across different systems.
- For low-resource languages, Unicode provides:
  - Standardization of character sets.
  - The ability to handle multiple languages in one document or system.
  - Compatibility with a wide range of NLP tools and resources.

Challenges in Unicode Processing

- **Encoding Issues:**
  - Variations in Unicode sequences for the same character.
  - Example: Hindi word "क्यों" has multiple valid Unicode sequences.
  - Solution: Normalization libraries like Python's `unicodedata` for consistent representation.

Challenges in Unicode Processing (Continued)

- **Missing Representations:**
  - Scripts like N'Ko or Vai lack adequate representation in Unicode.
  - Example: Vai script texts historically relied on custom fonts, leading to rendering issues.
  - Solution: Unicode expansion, such as Adlam script inclusion in Unicode 9.0.
- **Script Ambiguity:**
  - Shared scripts (e.g., Cyrillic for Serbian and Russian) complicate accurate language identification.
  - Example: Cyrillic letters like "а" and "е" appear identical in multiple languages but have distinct linguistic roles.
  - Solution: GlotScript assists in script and language disambiguation.
- **Non-Standardized Punctuation:**
  - Multiple Unicode representations for symbols like quotation marks.
  - Example: Double quotes may appear as 'U+0022' or 'U+201C', leading to parsing errors.
  - Solution: Text cleaning tools like "ftfy" standardize punctuation.

Unicode Resources for Low-Resource Languages

Normalization Tools:

- Python's `unicodedata` and `ftfy` lirbraries: Standardizes Unicode encodings.
- ICU Project[1]: Provides libraries and tools for Unicode support and internationalization.

Script and Symbol Support:

- Google Noto Fonts[2]: Comprehensive script coverage.
- Unicode CLDR Project[3]: Locale-specific data for languages and symbols.
- Unicode Character Database (UCD)[4]: Repository for Unicode character properties.

Script Identification:

- GlotScript[5]: Identifies and disambiguates scripts across languages.

[1] https://icu.unicode.org
[2] https://fonts.google.com/noto
[3] https://cldr.unicode.org
[4] https://www.unicode.org/ucd/
[5] https://arxiv.org/abs/2309.13320

Creating a Wikipedia for low-resource languages

Benefits for Native Language Communities

- **Education:** Access to native-language content improves literacy.
- **Collaboration:** Encourages speakers to grow their language's digital presence.
- **Heritage:** Safeguards and shares cultural, historical, and linguistic knowledge for future generations.

Benefits for NLP Research

- **Data Creation:** Produces a valuable resource for tasks like machine translation and knowledge graphs.
- **Model Input:** Provides structured datasets for training NLP models.

Example Research Project: WikiTransfer[1]

Data

- Biographies in high-resource language (*English*) and corresponding low-resource language (*Hindi*) collected from Wikipedia

Process

1. **Section Mapping:** Identify similar sections between the two language articles using section headings.
2. **Translation:** Use machine translation to translate content from high-resource to low-resource language.
3. **Content Enrichment:** Add only the new, unique content from the translation to the low-resource article.

(1) WikiTransfer: Knowledge transfer from High-Resource to Low-Resource Language in Multilingual Wikipedia
Paramita Das, Amartya Roy, Animesh Mukherjee
Wiki Workshop (11th edition, 2024)

Chair for Data Analytics & Statistics
School of Computation, Information and Technology (CIT)
Technical University of Munich

Spell Checking, Tokenization, and Word Segmentation

Chair for Data Analytics & Statistics
TU München

Spell Checking, Tokenization, and Word Segmentation

- **Spell Checking**:
  - Correcting typographical errors in text, ensuring the spelling of words is accurate and standardized.
  - Critical for improving the quality of text data before processing with NLP models.
- **Tokenization**:
  - The process of splitting a stream of text into meaningful units like words or subword units.
  - Key for NLP tasks such as translation, named entity recognition, and sentiment analysis.
- **Word Segmentation**:
  - Particularly relevant for languages like Chinese, Thai, and Japanese, which don't use spaces between words.
  - Essential for accurate translation and other NLP tasks.

Resources and Tools

- Subword Tokenzation: Byte Pair Encoding[1] (BPE), Sentencepiece[2], ByT5[3]
- Spell Checking: Grammarly[4], QuillBot[5]
- Word Segmentation: Stanford Word Segmentor[6], T-LAB[7] (Chinese and Japanese)

[1] Neural Machine Translation of Rare Words with Subword Units (Sennrich et al., ACL 2016)

[2] SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing (Kudo & Richardson, EMNLP 2018)

[3] ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models (Xue et al., TACL 2022)

[4] `https://www.grammarly.com/`

[5] `https://quillbot.com`

[6] `https://nlp.stanford.edu/software/segmenter.shtml`

[7] `https://tlab.it/en/allegati/help_en_online/msegment.htm`

Part-of-Speech Tagging and Morphological Analysis

Part-of-Speech Tagging

- POS tagging: annotation of word classes (POS tags)
- Possible extension: annotation of morphological information

- Example (Finnish): *"Maybe the dress code was too stuffy."*

Input	**POS tags**	**morph. tags**
Ehkäpä	ADV	ADV
pukukoodi	NOUN	NOUN.Nom.Sg
oli	AUX	AUX.Fin.Ind.Sg.3.Past.Act
liian	ADV	ADV
vanhanaikainen	ADJ	ADJ.Nom.Pos.Sg
.	PUNCT	PUNCT

Part-of-Speech Tagging: Training and Data

- There are many options to train a tagger
- If (nearly) no data of the target language is available and target language is similar to a high(er) resource language
  - fine-tune POS tagging for a supporting language
  - annotated target-language data can be helpful (>100 sentences)
  - Paper: Vandenbulcke et al. (2024): *Recipe for Zero-shot POS Tagging: Is It Useful in Realistic Scenarios?* MRL 2924.
    `https://aclanthology.org/2024.mrl-1.9.pdf`
- Training a tagger from scratch
  - training data: text annotated with POS tags (>1k sentences ?)
  - Paper: Schmid (2019): *Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts.* DATeCH 2019.
    `https://www.cis.uni-muenchen.de/~schmid/papers/Datech2019.pdf`

Morphological Analysis and Generation

- Morph. analysis: decomposition into linguistically meaningful units

  `Apfelbaums` → `Apfel<NN>Baum<+NN><Masc><Gen><Sg>` (*'apple tree'*)

- Morph. generation: forming an inflected form given the stem and relevant features

  `Apfel<NN>Baum<+NN><Masc><Nom><Pl>` → `Apfelbäume` (*'apple trees'*)

- Features, depending on the language:
  - inflectional features: morpho-syntactic features (for example number, case, tense, ...)
  - derivational features and word-formation (for example nominalization and compounding)

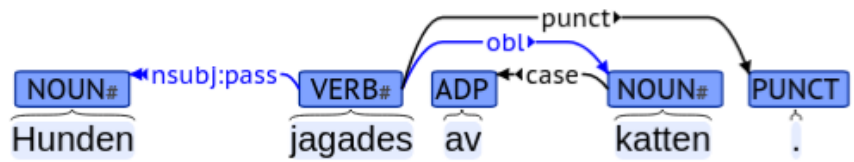
Morphological Analysis and Generation: Approaches

- Finite-state based morphological analysis/generation:
  - manual implementation of morphological rules
  - no training data needed, but linguistic expertise
  - Paper: Hulden (2009). *Foma: a Finite-State Compiler and Library*. EACL 2009. `https://aclanthology.org/E09-2008/`
- Neural morphological generation
  - Encoder-decoder model
  - Training data: pairs of feature - word pairs ($>$ 1000 examples ?)
  - Paper: Kann et al. (2016): *Single-Model Encoder-Decoder with Explicit Morphological Representation for Reinflection.* ACL 2016. `https://aclanthology.org/P16-2090.pdf`

Universal Dependency Treebank

Universal Dependency Treebank: Overview

- Cross-linguistically consistent treebank annotation for many languages
- Facilitate multilingual language research (for example parser development and cross-lingual learning)
- Example (Swedish): *The dog was chased by the cat.*



1	Hunden	hund	NOUN	_	Definite=Def	2	nsubj:pass	_	_
2	jagades	jaga	VERB	_	Tense=Past|Voice=Pass	0	root	_	_
3	av	av	ADP	_	_	4	case	_	_
4	katten	katt	NOUN	_	Definite=Def	2	obl	_	_
5	.	.	PUNCT	_	_	2	punct	_	_

UD: Format and Adding New Languages

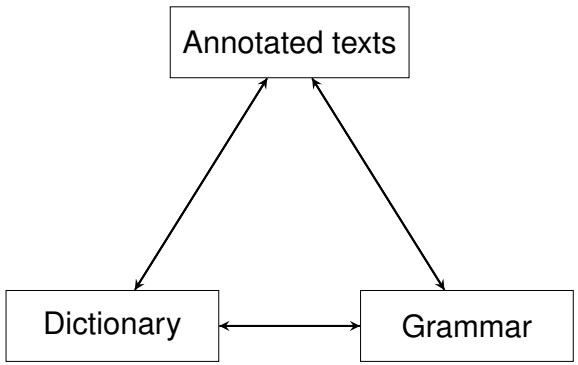
- A treebank entry contains manually revised dependency-parsed data
- CoNLL-U format: tab-separated fields to list a word's features and dependencies
- Add a new language:
  - select a data set that has no copyright issues (e.g. 1000 sentences)
  - pre-processing: there might already exist tools for your language or a related language
  - annotation: several existing annotation and visualization tools
  - annotator agreement: ideally, a new treebank is annotated by a team
- Further reading and a list of resources:

  `https://universaldependencies.org/how_to_start.html`
- Paper: Marneffe et al. (2021): *Universal Dependencies.* Computational Linguistics, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.

  `https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies`

Language Documentation

- Language documentation: field of linguistics which focuses on creating and archiving resources on (mostly) endangered languages
- Related to (but different from) preservation or revitalisation, which both focus on the use of the language



Boasian trilogy

Main challenge: annotation bottleneck

- Requires significant linguistic expertise and is hence time-consuming
- e.g., transcribing one hour of audio recording can take from 40 to 100 hours (Seifart et al., 2018)

Computational Language Documentation (CLD)

- Computational Language Documentation aims to create tools and models to assist **linguists** in their documentation tasks by automating certain steps as much as possible
- Linguistic annotations represent analyses, which require human control and verification

S0	Audio recording	Audio file		
S1	Unsegmented transcription	kɤndʑixtɤɣχɯmpjɤtunɯ		
S2	Segmented into words	kɤndʑixtɤɣ	χɯm	pjɤtunɯ
S3	Segmented into words and morphemes	kɤndʑi–xtɤɣ	χɯm	pjɤ–tu–nɯ
S4	Glossed sentence	COLL–brother	three	IFR.IPFV–exist–PL
S5	Translation (EN)	*There were three brothers.*		

Main annotation tiers of a sentence in (computational) language documentation (in Japhug)

Example of CLD tasks:

- Transcription (phonetic or almost phonetic)
- Sequence segmentation: word segmentation, morpheme segmentation
- Automatic generation of glosses (linguistic annotations)

Articles on Computational Language Documentation: (Zariquiey et al., 2022), (Gessler, 2022)

Chair for Data Analytics & Statistics
School of Computation, Information and Technology (CIT)
Technical University of Munich

Applied NLP tasks

Chair for Data Analytics & Statistics
TU München

Applied NLP Tasks for Low-Resource Languages

- **Examples of Applied Tasks:**
  - Hate Speech Detection, Fake News Detection, Text Summarization, Entity Recognition, and etc.
- **Challenges:**
  - **Data Scarcity:** Annotated datasets are limited for many languages.
  - **Linguistic Diversity:** Dialects, code-switching, morphology, g agglutination, and tonality.
  - **Annotation Costs:** High cost of recruiting native speakers and domain experts.
  - **Domain Adaptation:** Lack of domain-specific data limits cross-domain performance.
  - **Evaluation Gaps:** Few standardized benchmarks for low-resource languages.
- **Data Requirements:**
  - Hate Speech Detection: 10k–50k examples
  - Fake News Detection: 5k–20k examples
  - Few-shot learning can yield moderate results with 100–500 annotated samples, but lacks robustness.

Advances, Tools, and Resources

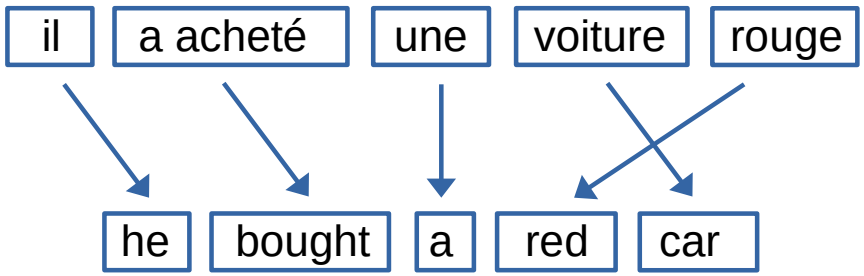
- **Recent Advances:**
  - **Data Augmentation:** Techniques like back-translation, paraphrasing, synthetic data.[1],[2]
  - **Active Learning:** Focused annotation on high-uncertainty samples.[3]
  - **Cross-Lingual Transfer:** Leveraging data from high-resource languages.[4]
- **Annotation and Inclusion Tools:**
  - Prodigy, Doccano, Label Studio, WebAnno

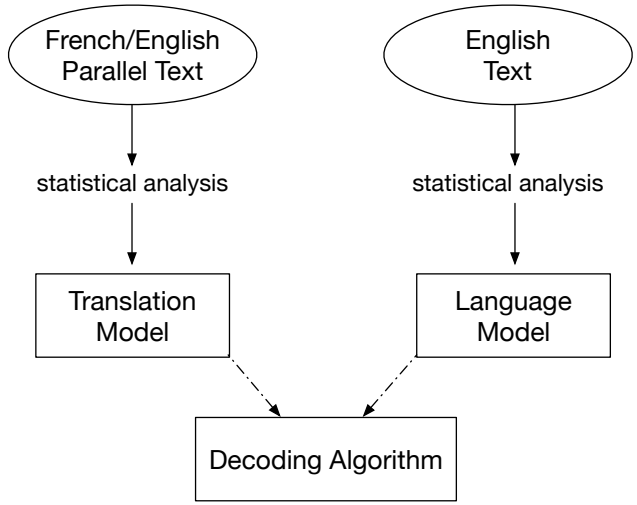
[1] Conneau et al., 2020, "Unsupervised Cross-Lingual Representation Learning at Scale." ACL.
[2] Liu et al., 2021, "Augmenting Low-Resource NLP with Synthetic Data." EMNLP.
[3] Siddhant et al., 2022, "Active Learning for Low-Resource Text Classification." NAACL.
[4] Artetxe et al., 2019, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer." ACL.

Statistical Machine Translation

Statistical Machine Translation: Idea

- SMT models have two main components:
  - **Translation model** to reproduce the source-side content: translation probabilities estimated from word-aligned parallel data
  - **Language model** to make the output fluent in the target language





- Decoding: the source input is segmented into phrases, each phrase is translated into the target language
- Phrases can be reordered
- Phrases do not necessarily correspond to linguistic units

SMT: Training and Resources

- Data required to train an SMT model
  - parallel data ($\geq$ 10k parallel sentences)
  - monolingual data for the target-side language model ($\geq$ 10k sentences)
- Pre-processing steps:
  - corpus cleaning, word tokenization
  - parallel data: word alignment
- Further reading:
  - `http://www2.statmt.org/moses/?n=Moses.Background`
  - Paper discussing the development of bilingual SMT models for translating English into fifteen low-resource Indic languages:

    Das et al. (2023): *Statistical Machine Translation for Indic Languages*
    `https://arxiv.org/abs/2301.00539`

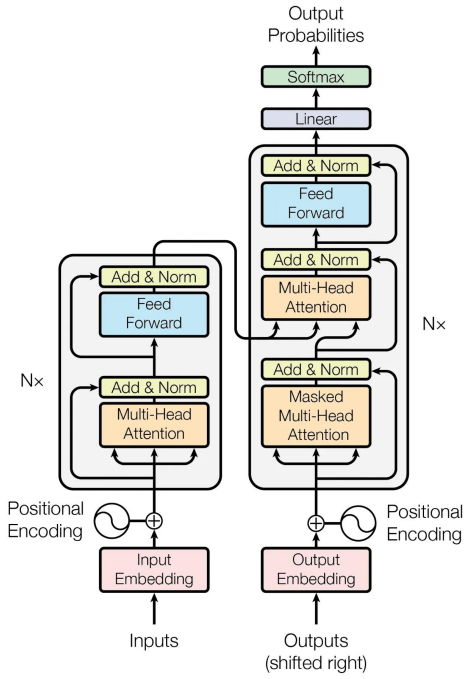
Chair for Data Analytics & Statistics
School of Computation, Information and Technology (CIT)
Technical University of Munich

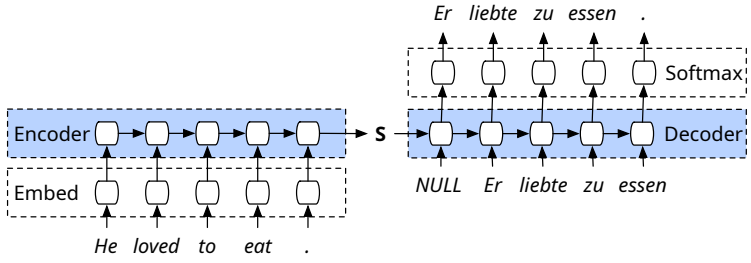
Neural Machine Translation

Chair for Data Analytics & Statistics
TU München

Neural Machine Translation: Idea

- NMT models rely on *deep learning* techniques to map source sentences to target sentences directly:
  - **Encoder-Decoder Architecture:**
    The encoder processes the source sentence and encodes it into a fixed-length vector representation (context vector). The decoder generates the target sentence based on this context.
  - **Attention Mechanism:**
    The attention mechanism allows the model to focus on different parts of the source sentence while generating each word of the target sentence.





NMT: Training and Resources

- Data required to train an NMT model
  - parallel data ($\geq$ 100K parallel sentences, though smaller datasets can also work)
  - optional: monolingual data for pre-training or fine-tuning (e.g., language models)
- NMT Tools:
  - `Fairseq`, `OpenNMT`, and `HuggingFace Transformers` are popular frameworks for implementing NMT.
- Data Resources:
  - `OPUS`, `NLLB`, `UN`, `Europarl`
- Further reading:
  - Survey of Low-Resource Machine Translation (Haddow et al., CL 2022)

Multilingual Chatbots

Continued pretraining of an English-centric model

Example 1: Adding Amharic to Llama2 (Andersland 2024)[1]

- 400M tokens of natural Amharic data
- Roughly 3.3B tokens of English data translated to Amharic

Example 2: Adding Tamil into Llama2 (Balachandran 2023)[2]

- 500M tokens of natural Tamil data

Lower resource languages: We don't know, it depends

- You can take advantage of related higher-resource languages
- If none are available, probably 400M as in examples 1 and 2

[1] https://arxiv.org/abs/2403.06354
[2] https://arxiv.org/abs/2311.05845

Instruction Fine-Tuning for Known Languages

This is needed to get good chatbot responses

Example 1: Adding Tamil into Llama2 (Balachandran 2023)[3]

- 140k examples of instructions translated from English

Example 2: Multilingual Model (Üstün et al. 2024)[4]

- As little as 400 examples for low-resource languages/dialects with high-resource relatives (e.g. Swiss German)

Recommendation: at least 100k examples

[3] https://arxiv.org/abs/2311.05845
[4] https://arxiv.org/abs/2402.07827

Training Chatbots from Scratch

Training from scratch

Pros:

- No bias from English or other languages in the model
- Can make the model smaller and therefore more efficient (for inference)

Cons:

- Need more GPUs for training
- Poorer quality output
- More steps to get a working chatbot

Training Chatbots from Scratch

BabyLM Challenge[1]

A challenge to train a model from scratch with very little English data, roughly the same amount a 13-year old has seen in their lives (100M words).

Results from the challenge:

- Some understanding of grammaticality (≈80% acc. across a wide range of grammatical phenomena)
- Some understanding of semantic relatedness (≈75% acc. for classifying entailment/neutral/contradictions)
- Poor understanding of world knowledge (e.g. A is left of B implies B is right of A)

[1] https://arxiv.org/abs/2412.05149

Training Chatbots from Scratch

Challenges:

- Scarcity of High-Quality and Digital Data
- Lack of Annotated Resources
- Linguistic Diversity and Complexity

Data Requirements for LLMs:

- Small Models ($< 1B$ parameters): Tens of millions of tokens; limited generalization.
- Medium Models (1–$10B$ parameters): $100B$–$1T$ tokens for optimal performance.
- Large Models ($> 100B$ parameters): Trillions of tokens; ideally 1 token per parameter.

Rule of Thumb: Tokens required $\approx$ Parameter Count $\times 20$ for general-purpose pretraining. (Chinchilla[2] paper)

[2] https://arxiv.org/abs/2203.15556

Training Chatbots from Scratch

Some Recent Solutions:

- Multilingual Vocabulary Optimization (SentencePiece[3])
- Efficient Pretraining Architectures (Switch Transformers[4])
- Adaptive Pretraining Objectives (CANINE: Token-Free Models[5])
- Curriculum Learning (Curriculum Learning for LLMs[6])
- Cross-Language Gradient Alignment (Gradient Alignment in Multilingual[7])

[3] https://aclanthology.org/D18-2012/
[4] https://arxiv.org/abs/2101.03961
[5] https://aclanthology.org/2022.tacl-1.5.pdf
[6] https://arxiv.org/html/2405.07490v1
[7] https://arxiv.org/abs/2109.08259

Speech Recognition

Advantages of Speech Recognition

- **Hands-Free Interaction:** Control devices without typing, enabling multitasking (e.g., driving).
- **Increased Accessibility:** Better access for people with disabilities.
- **Speed and Efficiency:** Faster than typing, improving productivity.

Challenges for Low-Resource Languages

- **Data Scarcity:** Limited annotated speech data for training.
- **Computational Needs:** High resources required for accurate models.

Speech Recognition

Data

- Paired text and speech data.

Process

- **Pre-Training**: Train model on large high-resource language dataset.
- **Fine-Tuning**: Adapt model with low-resource language data.

LRSpeech[1]

- Text Synthesis and Recognition system.
- Designed for (extremely) low-resource languages.
- Achieves promising accuracy with just hours of paired data.

[1] https://speechresearch.github.io/lrspeech/

Conclusion

- We hope the two parts of the workshop have been useful to you
- We also hope that this workshop will be the first of several zoom workshops
- We would like to create a Google Group for further communication, we'll send out the information using the registration form
- The slides will also be available

- Thank you for your attention!