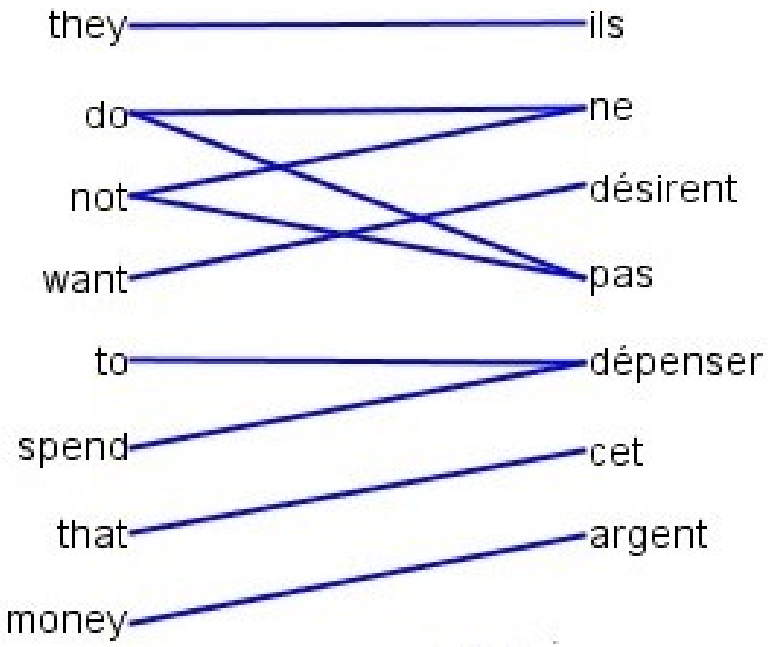
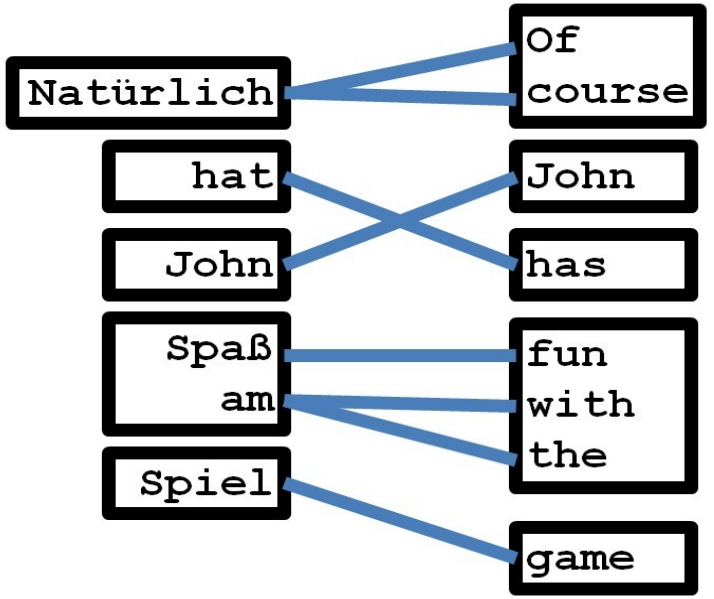




Some Open Problems in Multilingual NLP

Alexander Fraser

Center for Information and Language Processing

LMU Munich

Alexander Fraser

- I'm American, from Boston
  - However, despite this, I speak 4 European languages and Arabic

- PhD in Computer Science
  - 2007 University of Southern California / Information Sciences Institute
  - Work in Intelligent Systems Division (AI department: Daniel Marcu, Kevin Knight, Ed Hovy)
  - Also extensive industry experience (first statistical machine translation product) and additional international non-profit experience

- Since 2007 in Germany
  - Currently Professor of Information and Language Processing at LMU Munich
  - Tenured as of September 1st, 2020. In both Lang/Lit and CS/Math/Stat faculties
  - Interdisciplinary teaching: mainly Machine Learning (ML) and Natural Language Processing (mostly ML approaches), but also some linguistics and basic computer science/programming
  - Masters coordinator

Motivation: Machine Translation

- How can we break through language barriers?
- How can we …

  … find all of the information there is on a topic on the web, no matter what language it is written in?

  … understand newspapers around the world?

  … translate things that otherwise would not be translated at all due to manpower/financial constraints?

  … automate boring repetitive translation tasks, allowing human translators to focus on fun and challenging translations?

  … create content in minority (low resource) languages?
- Solution: high quality machine translation!

Data-Driven Machine Translation

- Previous approach was so-called rule-based machine translation
  - Human experts writing rules
- Current state-of-the-art uses supervised machine learning: learn how to translate from examples
  - Examples are pairs of sentences (a sentence and its translation)
- Phrase-Based Statistical Machine Translation (PBSMT), previously best, still used in some scenarios
- Neural Machine Translation (NMT), deep learning approach

Why is data-driven MT research interesting?

- Structured prediction
  - Sentiment Analysis is not structured prediction: label a movie review with one of 3 classes: positive, neutral or negative sentiment
  - Machine Translation is structured prediction: label a 30 word English input sentence with a 28 word German translation (!)
- Uses world and contextual knowledge (later in talk)
- Evaluation
  - There are many right answers, the training data contains just one of the alternatives!
- Applicability
  - MT is basically a language modeling problem. Anything with text outputs is also a language modeling problem.
  - Feature engineering on text is done with representations from language models and MT (e.g., ULMFiT, BERT, MASS, …). Our research: multilingual representations
  - We can apply MT models to problems like image captioning with little change, just combine an image encoder with our standard text decoder

What I'll Talk About

- The talk will be in two parts
  - In the first part, I'll give you a brief idea of some research we have done on domain adaptation
  - In the second part, I'll complain about (multilingual) NLP research. Some other things we really should address:
    - Multimodality
    - Callibration
    - It's the training data, stupid!
    - Clever Hans
    - Explaining Explainability
    - User modeling

Domain adaptation for MT

- MT works well when translating sentences from the same domain as the parallel training data
- **What about new domains**?
- In domains like consumer health or medical, we have little or no parallel data
  - How can we deal with this problem?
- I organized a "summer workshop" (= crash research project, 13 people for 6 weeks) at Johns Hopkins on this topic
  - Co-organizers: Hal Daume (Maryland), Marine Carpuat (National Research Council Canada), Chris Quirk (Microsoft Research)
- I was awarded an **ERC Starting Grant** by PE6 (Computer Science) to continue this work and try a number of new approaches to solve this problem
  - I will present our work on "Document as Domain" in some detail

ERC StG: Domain Adaptation for MT

- My ERC is on Domain Adaptation for MT
- Traditional domain adaptation techniques in SMT and NMT have focused on the corpus as a proxy for domain
- If we have plentiful parallel data in the legal domain, we can translate legal documents
- But what if we do not have such data?

Roadmap: Domain Adaptation

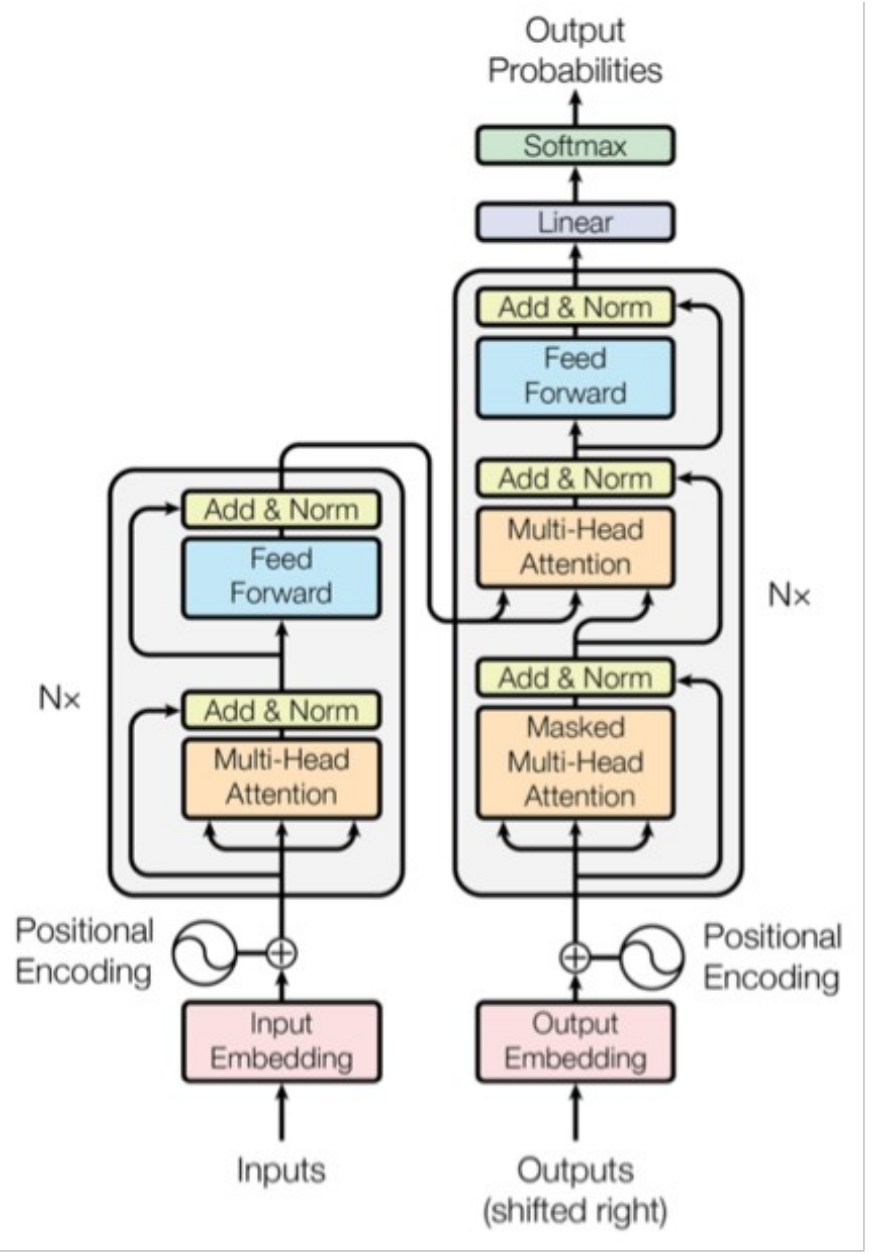
- I will first briefly introduce NMT in detail
- Then I will contrast three approaches to domain adaptation
- The running example is the translation of this English snippet to German

Input: … that is a beautiful **seal**

- But first some basics (not our work!)

Transformer NMT - Encoder

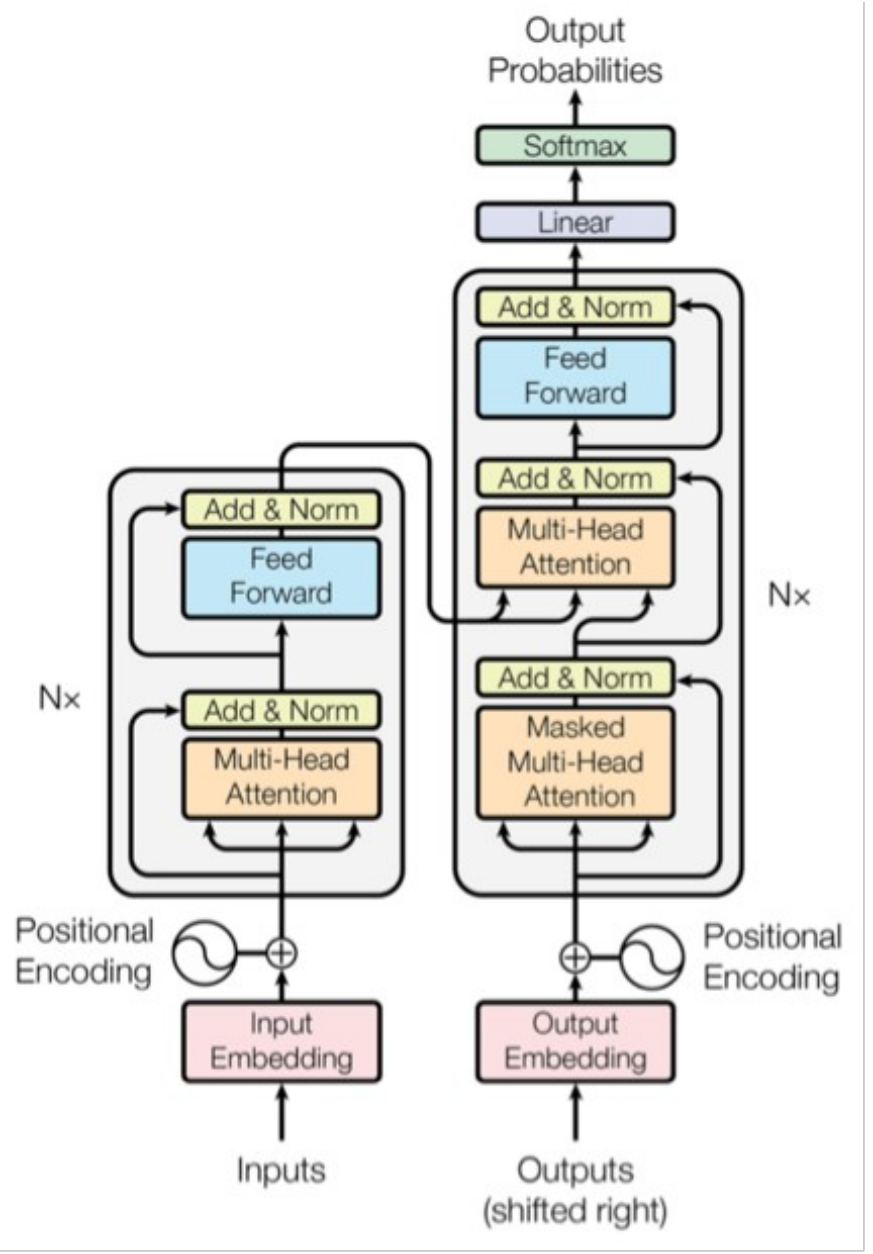
- The Transformer is the state-of-the-art sentence level model for NMT
  - This has largely replaced Recurrent Neural Network based formulations
- The lefthand side shows the encoder
- Inputs are a sequence of words
- The first layer is a set of word embeddings (one per word-type)
- This input is processed using 6 layers of feed forward networks with attention
- Attention allows the network to focus on what is important for each position



Graphic from Vaswani et al. 2017, p 3

Transformer NMT - Decoder

- The righthand side shows the decoder
- The decoder receives as input first a start signal and then the decoder outputs shifted right by one timestep
- This is also processed using 6 layers of feed forward networks with attention to the input
- But there is additional Masked Self-Attention
  - Self-Attention allows the decoder to give attention to previously output positions
  - Masking blocks it from looking at the current or future positions during training



Graphic from Vaswani et al. 2017, p 3

No domain knowledge

- … that is a beautiful seal .
- … das ist ein schöner Seehund. (animal sense)
- Looks great?
- Here is some context: I asked the notary. She said that is a beautiful seal.
  - Try this in Google Translate – it gets seal right! (checked again earlier today)
- Different context: I asked the zookeeper. She said that is a beautiful seal.
  - Try this in Google Translate – it gets seal wrong!

How to model domain?

- Just add an additional domain marker to the source language sentences (Kobus et al. 2017)
  - This marks source sentences with the corpus they came from
- Then retrain the transformer
- When translating: provide the domain marker for future sentences

Input: **<LEGAL>** I asked the notary. She said that is a beautiful seal.
Output: … das ist ein schönes **Siegel.**

Input: **<GENERAL>** I asked the zookeeper. She said that is a beautiful seal.
Output: … das ist ein schöner **Seehund.**

Problems with domain tags

Cool, problem solved!

Input: **<PLUMBING>** I asked the plumber. She said that is a beautiful seal.

Wait, where do I get parallel data for the plumbing domain?

Also, who is giving me the <PLUMBING> tag, I don’t see where to put this in Google Translate?

The answer btw: **Dichtung**

Document as Domain

- People try to solve this using classifiers (usually on the input sentence)
  - But this relies on explicit domains at the corpus level
- We do not believe in corpus-level domains
- Instead, we build **document-level NMT models**
- Most state-of-the-art MT systems translate sentence by sentence
  - This is obviously wrong!
  - Input: I asked the notary. **She** said …
  - Output: I habe **den Notar** gefragt. **Sie** sagte …
  - Should be: **die Notarin**

Document-level Domain Adaptation for NMT

- We would like to condition the translation of all words on their document-level context
- The baseline model does this very well for single sentences
  - However, attention is quadratic in the sentence length. We can't view a document as a long sentence!
- We have existing work on pronoun translation:

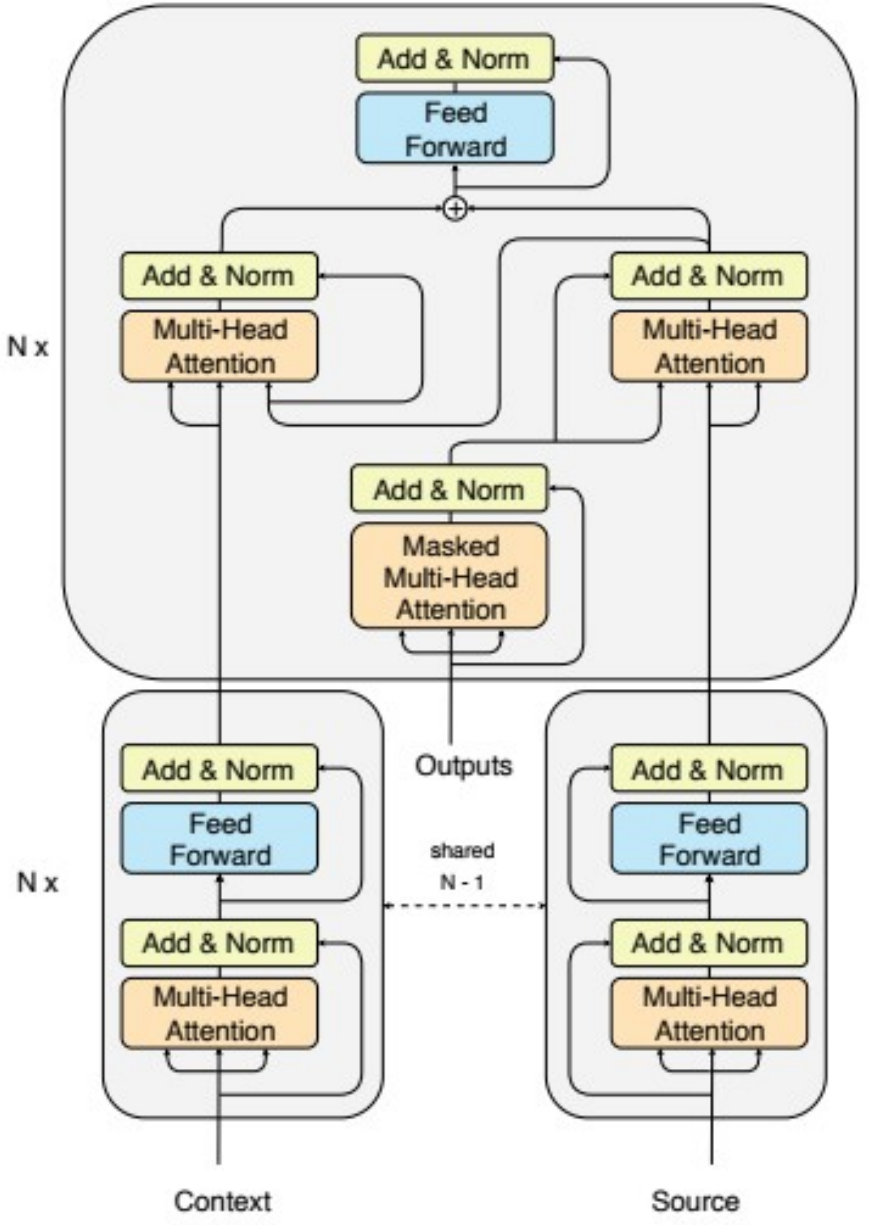
  Input: That is a beautiful dog. **It** ran away.

  Output: **... Er ...**

- New idea: model domain at the document level

Domain Adaptation Without Knowing the Domain

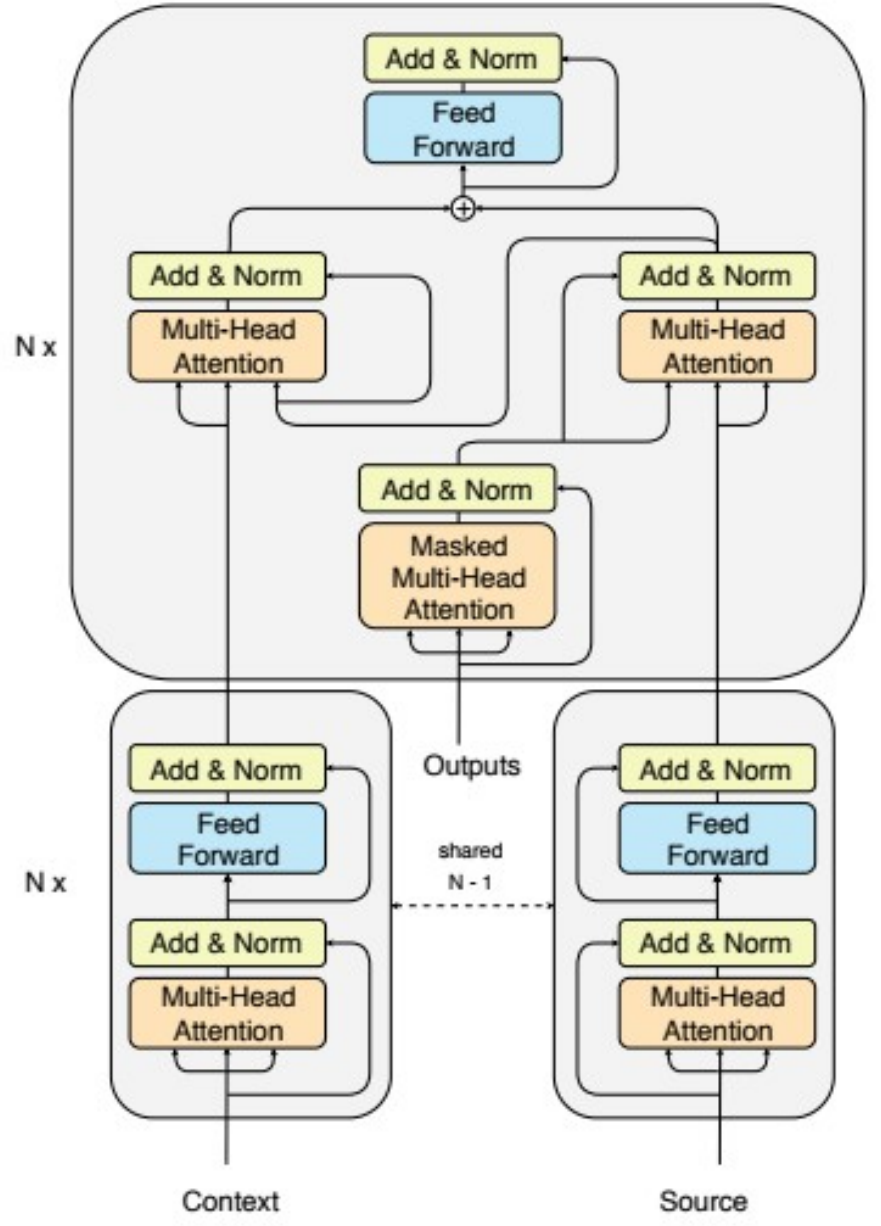
- We work with two models here, I will present these on the next slide
- The encoder shown to the right is from our Document NMT model, which we originally proposed for pronoun translation in 2019
- The part on the right is almost a standard Transformer encoder
- The part on the left encodes the context (context: the sentences in the document that we are currently not translating)
- The first 5 layers are shared
- The two representations are combined using a gate
- (There is also a decoder version of this, not presented)



Stojanovski and Fraser 2019, p 2

Domain Adaptation Without Knowing the Domain

- First model:
  - At the word level, add a document embedding
  - This is part of the input embedding
  - This is motivated by Kobus's domain tags, but we learn this end-to-end (like the embedding layer)
  - We use no knowledge of domain/corpus
- Second model (not shown):
  - Create a summarized representation of the document using max pooling over windows of 10 words for all context sentences
    - This effectively combines the contextual word embeddings
  - Also trained end-to-end, also no knowledge of domain/corpus



Stojanovski and Fraser 2019, p 2

Document as Domain - Results

- Summary of the results:
  - This approach is more powerful than previous work
  - Particularly strong when there is no training data for the domain
  - Even when the training data contains the domain, the baseline is given access to, e.g., <LEGAL> at both training and testing time, we are still somewhat better
    - We have no explicit knowledge of this (domain/corpus) information!
  - Also important: the domain embedding approach (first model presented) is also nearly as fast in decoding as the baseline, and it is resource efficient (see Stojanovski and Fraser 2020 for a comparison)
  - DeepL has recently started to translate some of my examples correctly (but not “den Notar. Sie …”).
    - I assume they are using a lightweight document encoder like the one I presented, implemented as a part of OpenNMT.

Two other projects

- Multilingual hate speech detection
- Moral language models

A few more slides

- Multimodality
- Callibration
- It’s the training data, stupid!
- Clever Hans
- Explaining Explainability
- User modeling

Multimodality

- It is time for NLP to move beyond text. Speech is the next obvious area to work on, but image processing is also not that hard anymore
  - Even image processing is using Transformers these days
- It is in fact likely that we can do a better job on text if we can leverage speech and image models
  - I'm interested here in multimodal detection of hate speech particularly
- Many early attempts at this seem to switch back and forth between two modalities, rather than jointly modeling them

Callibration

- Callibration is an elephant in the room for deep learning systems
- They are often overconfident, assigning a huge posterior probability to the answer that is selected
- For MT, the problem formulation isn't even right
  - Consider: "I saw the man with the telescope" translated into Chinese (which requires disambiguation). The posteriors don't look right!
  - Worse: there are many ways to correctly translate!!!
- We typically use a separate classification model to try to estimate the human evaluation score that will be given, primarily by using n-grams statistics on the parallel training corpora
  - This is ugly!

It’s the training data, stupid!

- Academic research holds the training data constant, and varies the model
- But everyone who has ever worked on a commercial system knows:
  - It’s the training data that matters!
  - There needs to be more work on this
- At the moment people see how far they can get with self-supervised pretraining like BERT (or mBERT, XLM-R, etc.)
  - This is actually pretty interesting, you can reduce annotated data for classification
  - But parallel data rules MT (at least currently). This is why DeepL beats Google at English to German translation.

Clever Hans

- I’ve made some progress on a lot of problems that I thought were quite difficult to solve
- One problem that I actually believed was solved (kind of embarassing) is pronoun translation for English to German
- There is a nice challenge set for this called ContraPRO, and Microsoft was getting very high scores on this
- In fact, it turns out their system was using superficial heuristics (simple statistics on the training corpora)
- We adversarially attacked ContraPRO and created ContraCAT. It was easy to show that getting a good score on ContraPRO was just Clever Hans (Linzen)

Explaining Explainability

- Explainability is just crazy difficult with these models
- I started working on MT just as it shifted from rule-based to statistical
  - Ironically, we initially thought statistical models couldn't be explained. Rule-based systems are easy to understand.
  - But we actually became adept at understanding decisions
- There is a lot to be done in understanding our deep learning models
- But there is also a lot to be done in evaluation of explanations!
  - How can we make progress is we can't evaluate? (Ideally give me an automatic metric that ranges between 0 and 100…)
  - Is "explanation" even really defined?

User Modeling

- Current academic MT systems take a sentence (or document) as input, and output a sentence (or document)
- But this isn't how people use MT!
- The problem is even worse for Multilingual NLP
  - Consider automatically detecting hate speech.
    - Building classifiers is hard, all of the previous problems apply
    - But even if we can build somewhat decent classifiers…
    - What does the user actually want? How will they use the classification decision? How can they understand the classificaiton decision?
    - How can we start to address this?

What I Talked About

- Mismatch of Train and Test (Domain Adaptation)
- Multimodality
- Callibration
- It’s the training data, stupid!
- Clever Hans
- Explaining Explainability
- User modeling

Thank You!

- Thanks for your attention
- Credits to my entire team, thank you!
- Contact: fraser@cis.lmu.de
- (or see my webpage, also for current and former team members, all publications are available)