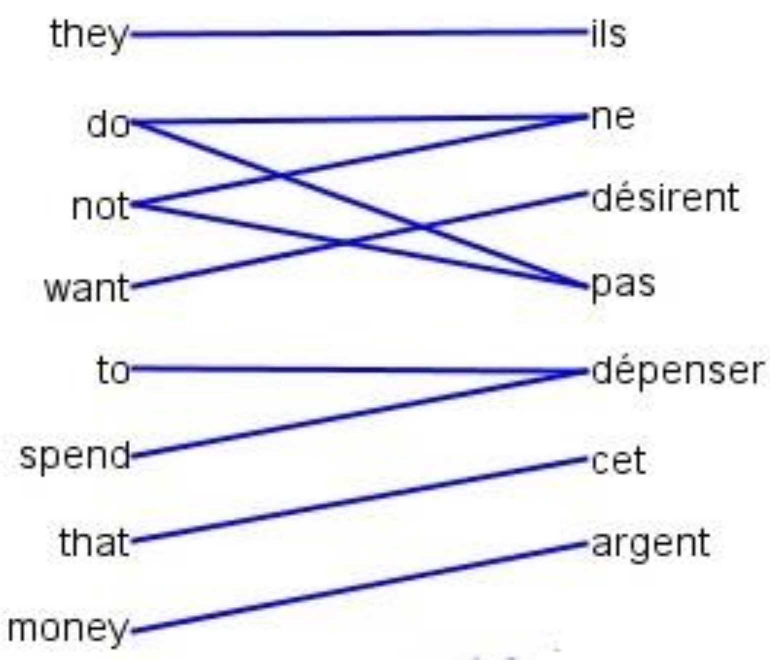
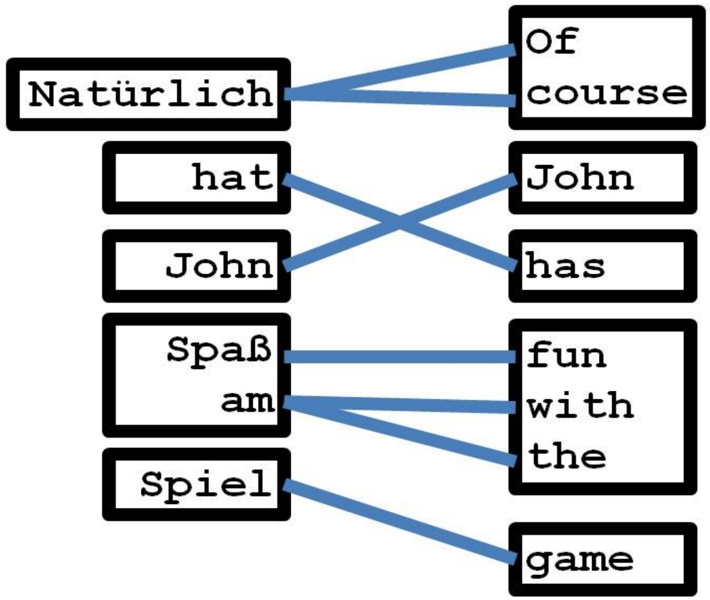




Overcoming Data Sparsity in Machine Translation

Alexander Fraser

Center for Information and Language Processing

LMU Munich

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?
- 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

3 weaknesses of data-driven MT

- I will talk about 3 problems with data-driven MT:
  1. Morphological richness causes data sparsity in MT
     - Solution: generalize over morphological phenomena
  2. MT is strongly dependent on the domain of the training data
     - Solution: develop new domain-adaptation techniques for MT
  3. MT is supervised, requiring a large number of parallel sentences
     - Solution: develop unsupervised MT

1. Morphological productivity: translating from English to German is difficult!

A huge problem in translating to German is **morphological productivity**

Words in noun phrases have context-dependent inflection in German

New German compounds are created every day!

DFG two-phase project and Health in My Language (HimL) Horizon2020 project

DFG project: four PhD theses (all four excellent female researchers)

Full English to German linguistic pipeline

- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, tense, aspect
- Translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features
  - I'll discuss this part briefly

Predicting nominal inflection

Idea: separate the translation into two steps:

(1) Build a translation system with non-inflected forms (lemmas)

(2) Inflect the output of the translation system

a) predict inflection features using a sequence classifier
b) generate inflected forms based on predicted features and lemmas

Example: baseline vs. two-step system

- A standard system using inflected forms needs to decide on one of the possible inflected forms:
  blue → blau, blaue, blauer, blaues, blauen, blauem
- A translation system built on lemmas, followed by inflection prediction and inflection generation:

  (1) blue → blau<ADJECTIVE>
  (2) blau<ADJECTIVE><nominative><feminine><singular> <weak-inflection> → blaue

Results and outlook

- Pipeline morphosyntax approach results in robust improvements in standard tasks (such as political/news)
  - See: Fraser et al. EACL 2012, Weller et al. ACL 2013, Cap et al. EACL 2014, four more *ACL papers
- Also improvements on medical translation tasks
  - HimL: EU Horizon 2020 Innovation Action with Edinburgh and Prague
  - Worked on consumer health information and deployed systems in two non-profits
  - Links up with my pre-PhD past. ICTs and health Information in developing countries at SatelLife
- No time to talk about:
  - We also implemented joint inference directly in the Moses PBSMT decoder (ACL and EACL papers)
  - Joint inference over morphology and word stems in Neural Machine Translation, large gains in performance on small to medium data sets (ACL and ACL WMT papers)
  - Related idea on linguistic segmentation won ACL Conference on MT shared task on English to German translation in 2017

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
  - Then I will quickly present the other of the two main areas we work in, mining from comparable corpora, which helps to setup "Unsupervised MT"

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
- 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**

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 earlier today)
- Different context: I asked the zookeeper. She said that is a beautiful seal.
  - Try this in Google Translate – it gets seal wrong (checked earlier today)

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:

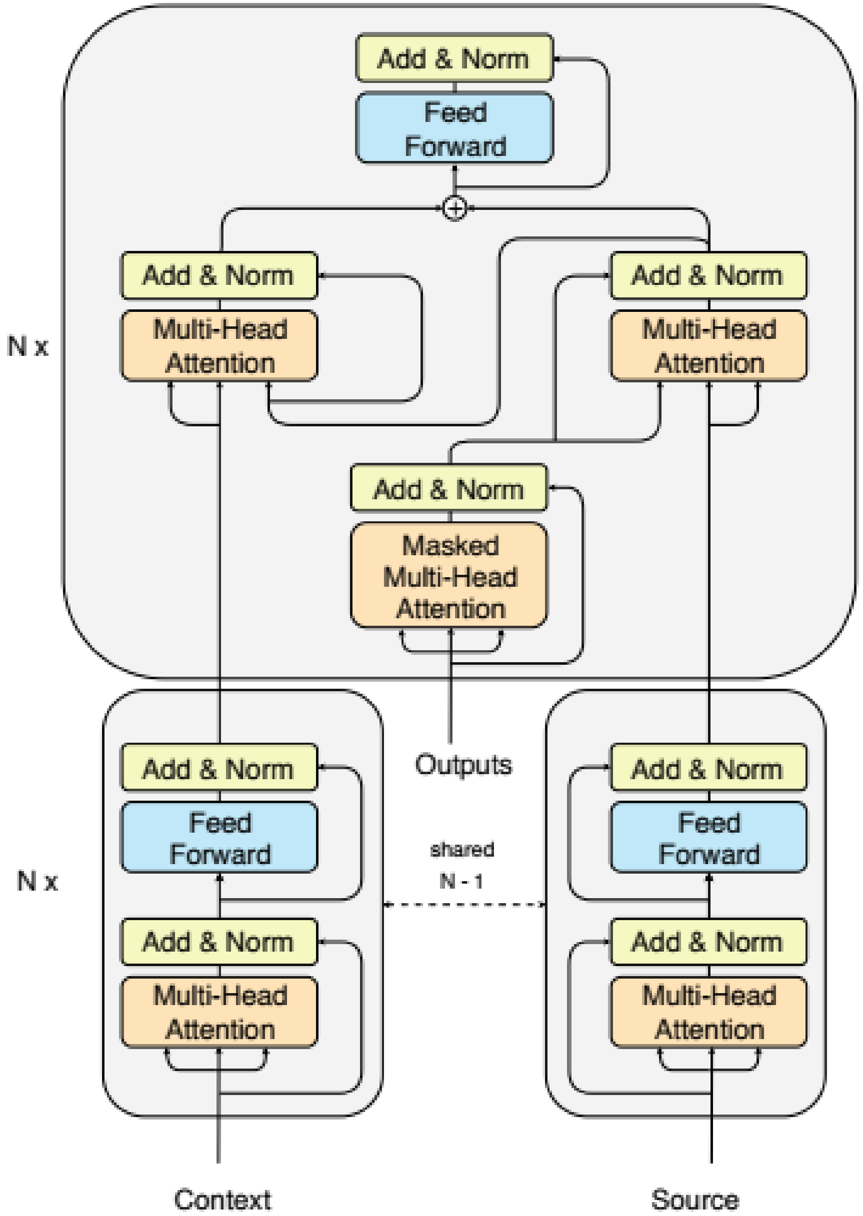
Input: That is a beautiful dog. **It** ran away.

Output: **… Er …**

- New idea: model domain at the document level

Domain Adaptation Without Knowing the Domain

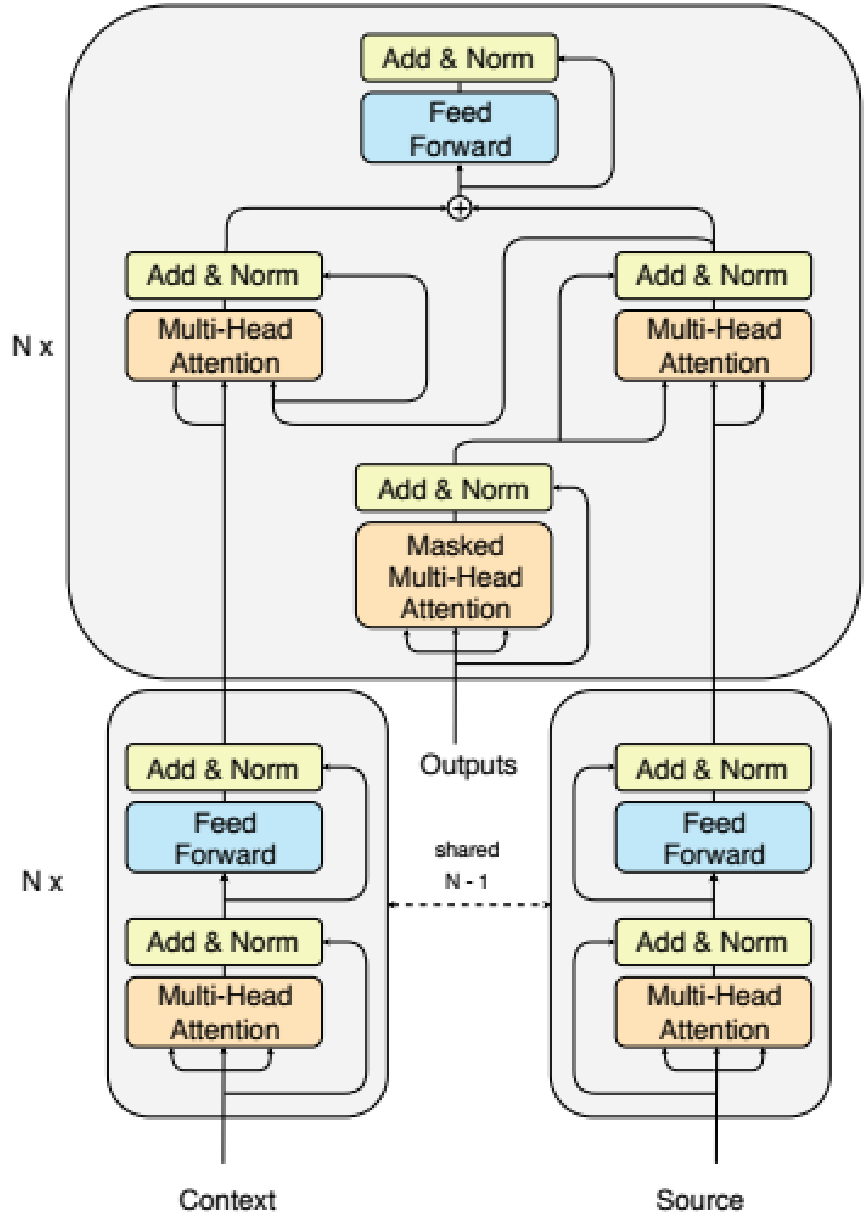
- We work with two models here, I will present these on the next slide
- The encoder shown to the right is from our Document NMT model, which we originally proposed for pronoun translation in 2019
- The part on the right is almost a standard Transformer encoder
- The part on the left encodes the context (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

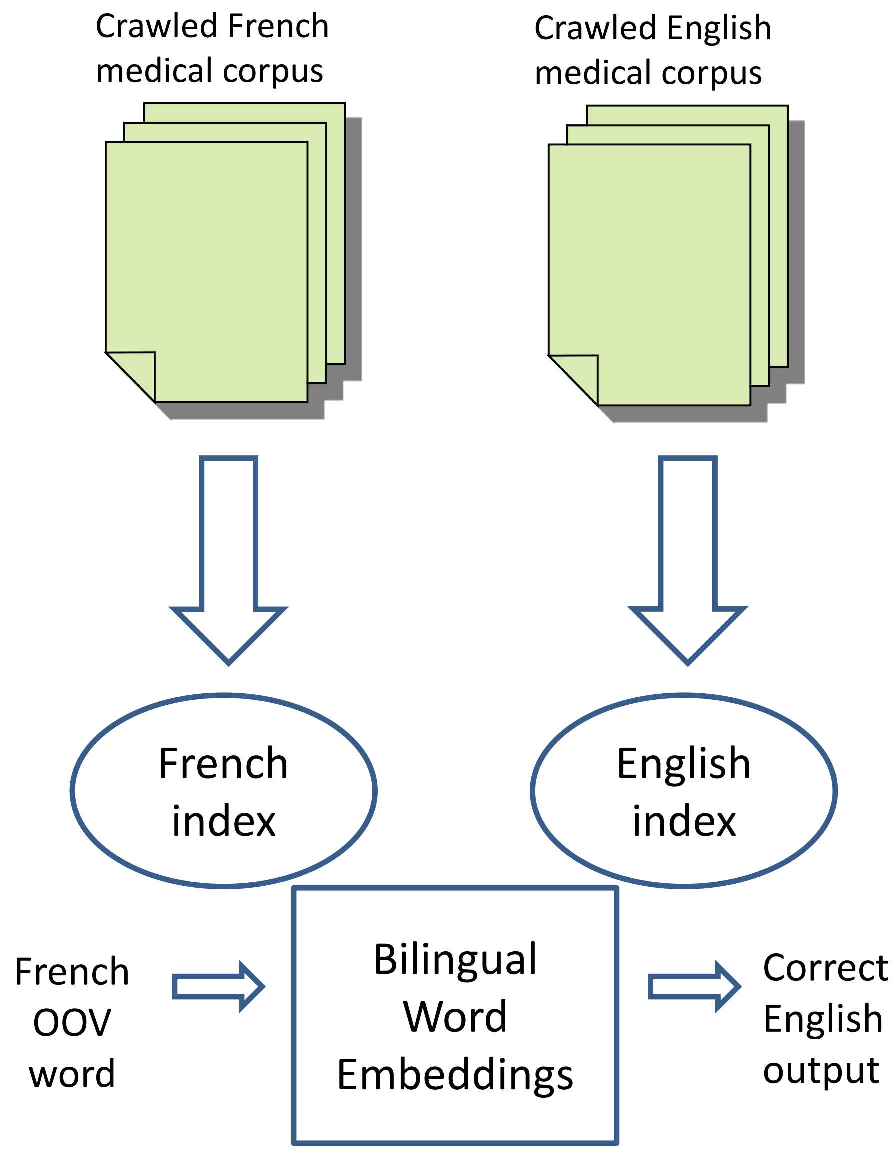
Document as Domain - Results

- 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)

- I will now present a few slides on comparable corpora, then unsupervised MT

Mining Comparable Corpora

- Key problem:
  - Crawl and index large corpora in the new domain
  - Find translations of new terms using bilingual word embeddings
- Our solution:
  - Use cutting edge retrieval techniques for crawling and indexing
  - Just in time terminology mining given new document to translate
    - Use context of source term as query to source corpora, then retrieve similar target language corpora and mine source term translation
    - Requires significant algorithmic innovation, previous approaches are offline and very slow!
  - Integrate terminology into morphology-aware machine translation to generalize to new contexts
  - Use novel domain fitness metric to score candidates, combine with document modeling



Results

- We have extensive work on mining translations that are semantically translated (e.g., Hangya, others, publications at NAACL, ACL)
  - Mostly lightly supervised systems
- We already have systems for unsupervised mining of:
  - Translations of words that are transliterated (Sajjad PhD thesis and several *ACL papers, work with Severini/Schütze)
  - Parallel sentences from comparable corpora (first paper on supervised extraction in 2004 with Munteanu, 2018 publications present first unsupervised extraction approach)
  - Out-of-vocabulary (OOV) words using word type Bilingual Word Embeddings (OOVs: words which we need to translate but that are not in our training data) (Huck et al ACL 2019)
  - Ongoing work with Anna Korhonen's group at Cambridge on using context-dependent embeddings

Segue to Unsupervised MT

- Since we know how to translate without using parallel corpora, we can begin to address low resource languages
- I'm enthusiastic about very low resource MT
  - Unsupervised MT is an interesting way to bootstrap a low resource language
  - Initially tried Upper Sorbian (next slides)
  - See also our Lower Sorbian task running now at ACL Conference on Machine Translation

Unsupervised Machine Translation I

- Wouldn’t it be great to translate to languages which don’t have large amounts of translations to any language available?
- For instance, minority languages
- Take the case of Upper Sorbian (spoken in the Oberlausitz in Eastern Germany)
  - Can we build MT from German to Upper Sorbian without parallel data?

Unsupervised Machine Translation II

- The answer is yes!
  - We can use unsupervised techniques to create Bilingual Word Embeddings between Upper Sorbian and German
  - Use these to generate rough translations of real Upper Sorbian sentence to (broken, noisy) German sentences
  - Then build an neural machine translation system that is robust to noise in the input to translate real German to real Upper Sorbian
  - Can iterate translation in both directions to further improve system

Unsupervised Machine Translation III

- We have fairly good systems already
- We are working on:
  - Improving bilingual word embeddings – see papers by Viktor Hangya
  - Increasing robustness to noise
  - Scaling of GPU training
  - Implementing semi-supervised pipelines
  - Pretraining using transformer language models (like mBERT) – see papers by Alexandra Chronopoulou
- I am the main organizer of the shared task on this at the ACL Conference on Machine Translation (2020, 2021)
  - In 2021 we will also study Russian-Chuvash and Lower Sorbian
- Finally: this also has important implications for building cross-lingual NLP systems, particularly for low resource languages

Summary

- I presented work in three areas:
  - Syntax and morphology for MT
  - Domain adaptation for MT
    - The focus in this talk was here, on using document models for domain adaptation
  - Unsupervised machine translation

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)