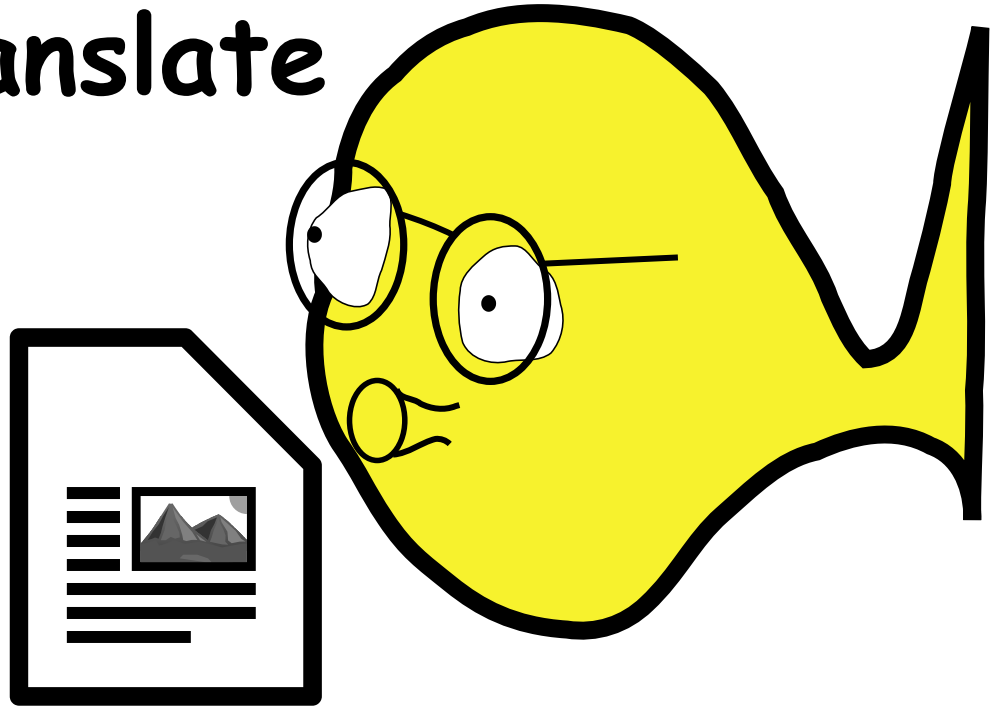


LibreOffice Translate

Offline AI-powered
Machine Translation
with a single Click



Built with: LibreOffice, PyTorch, OpenNMT

<https://github.com/lernapparat/lotranslate/>



LibreOffice
The Document Foundation



Neural Machine Translation for LibreOffice

Thomas Viehmann
tv@lernapparat.de

ALMERIA | 13 Sept. 2019



Agenda

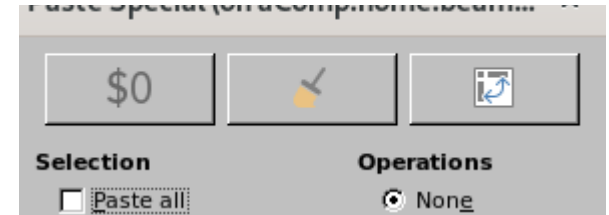
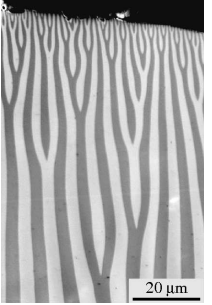
- About me
- LibreOffice Translate – An Extension for Writer
- How much UI do we need?
- Under the hood
 - Neural Machine Translation
 - NMT models meet the real world
 - PyTorch & OpenNMT
 - How to train new models
 - Pains of a LibreOffice Python extension with many dependencies
- Better Integration with Writer / other parts LibreOffice



Hi, I'm Thomas Viehmann

My background:

- did some Pattern Recognition and Neural Networks in 2000, but moved on to mathematics
- Ph.D. in Mathematics, modelling patterns in magnets
- 9 years of insurance risk modelling
- Recently lots of modelling in the ML context
- 2 tiny LibreOffice patches



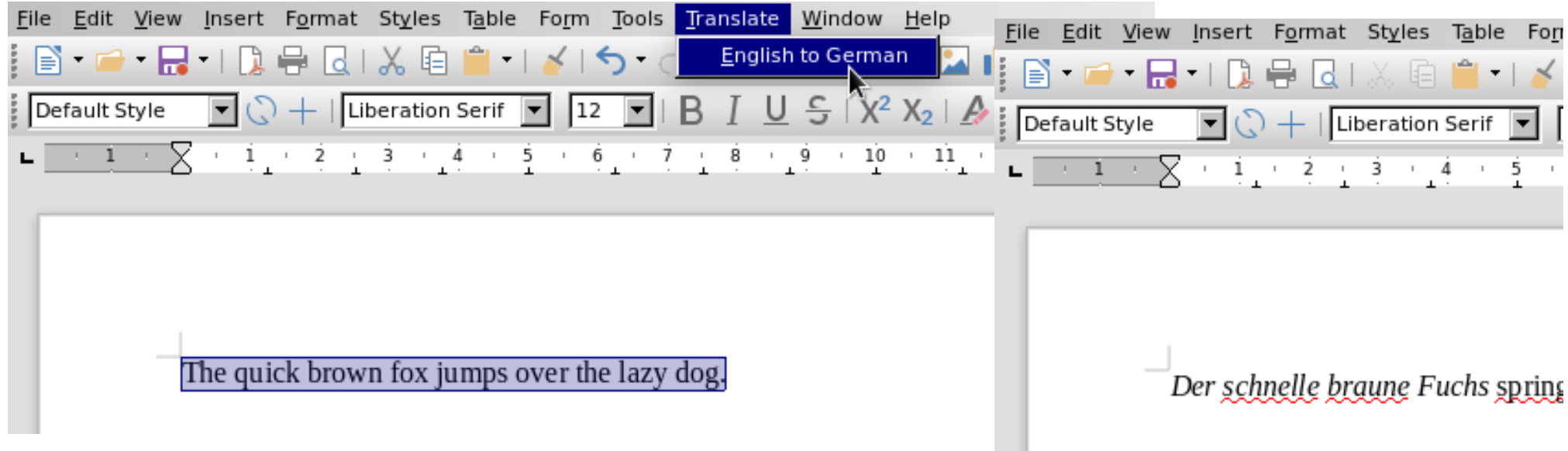
ML Training & Consultancy through: <http://mathinf.eu/>

Blog about ML stuff: <http://lernapparat.de/>



LibreOffice Translate

Idea: very simple (1-click) machine translation of text in Writer



<https://github.com/lernapparat/lotranslate/>

Development supported by the German Ministry of Research and Technology through the Prototype Fund. (Thank you!)



How much UI do we need?

Researched other Translation tools workflow:

“Text→Sidebar→Translation → Edit → Insert”

What is the advantage of using the sidebar?

→ You see original and translation side by side.

(Happily, the Prototype Fund sponsored a coaching regarding UI / UX which helped in reflecting this.)



How much UI do we need?

LOTranslate use Annotations instead (also saves UI coding):

The image displays two side-by-side screenshots of a web editor interface, likely LOTranslate, illustrating the use of annotations instead of UI coding.

Left Screenshot (Clean Document):

- Menschheitstraum**
Das [Verstehen](#) einer Sprache, ohne sie gelernt zu haben, ist ein alter Menschheitstraum ([Turmbau zu Babel](#), [J. Bechers numerische Interlingua](#), [Timerio](#), [Babelfisch](#), [Pfungstwunder](#), [Science-Fiction-Geschichten](#)). Die Erfindung der [Computer](#) in Kombination mit der Beschäftigung mit dem Phänomen Sprache als wissenschaftliche Disziplin ([Sprachwissenschaft](#)) hat zum ersten Mal einen konkreten Weg zur Erfüllung dieses Traums geöffnet.
- Statistische MÜ**
(*Statistics-Based Machine Translation, SBMT*)
Vor der eigentlichen Übersetzung analysiert ein Programm ein möglichst großes [Textkorpus](#) von zweisprachigen Texten (oft zum Beispiel Parlamentsprotokolle, etwa aus dem kanadischen Hansard-Corpus). Dabei werden Wörter und grammatische Formen in Ausgangs- und Zielsprache aufgrund ihrer Häufigkeit und gegenseitigen Nähe einander zugeordnet und somit ein Wörterbuch sowie Grammatikübertragungsregeln extrahiert. Auf dieser Basis werden die Texte übersetzt. Die statistische MÜ ist sehr populär, weil sie keinerlei Kenntnis der beteiligten Sprachen voraussetzt. Deshalb kann die statistische MÜ durch die Analyse realer Textbestände theoretisch auch solche Regeln erfassen, die sprachwissenschaftlich noch nicht genau erklärt sind.

Right Screenshot (Document with Annotations):

- Humanity dream**
[Understanding](#) a language without learning it is an ancient dream of mankind ([Turmbau zu Babel](#), [J. Becher's numerical Interlingua](#), [Timerio](#), [Babel Layer](#), [Pentagon Science](#)). The invention of computers in combination with the study of the phenomenon of language as a scientific discipline ([language science](#)) has for the first time opened up a concrete way to [fulfill](#) this dream.
- Statistische MÜ**
(*Statistics-Based Machine Translation, SBMT*)
Before the actual translation, a [programme](#) analyses as [large a text body](#) as possible of bilingual texts (often parliamentary minutes, for example, from Canadian Hansard Corpus). In doing so, words and grammatical forms in the source and target language are assigned to each other due to their frequency and mutual proximity, thus extracting a dictionary and grammar transfer rules. The texts are translated on this basis. The MET is very popular because it does not require any knowledge of the languages involved. Therefore, by analysing actual texts, the MÜ statistical study can theoretically also cover rules that have not yet been properly explained in terms of language.
- Neural MÜ**
(*Neural Machine Translation, NMT*)

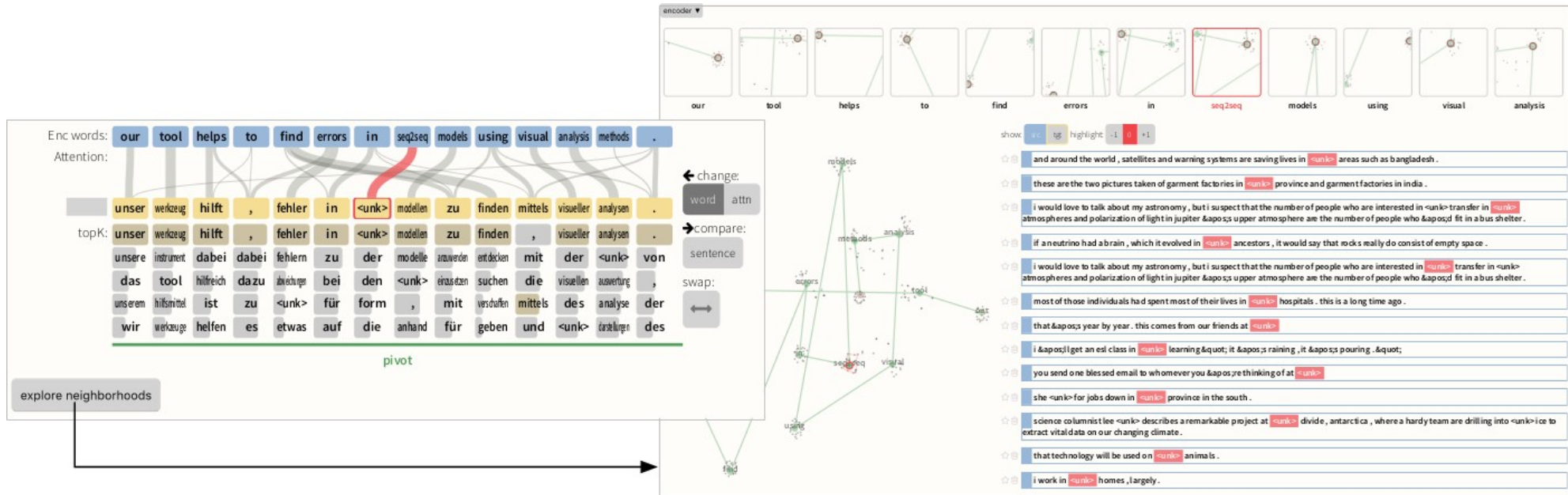
The right screenshot shows a sidebar with a list of LOTranslate annotations for each section, including dropdown menus and labels like "LOTranslate (no date)".

...annotations are optional



How much UI do we need?

There are interesting visualizations of model "innards"...

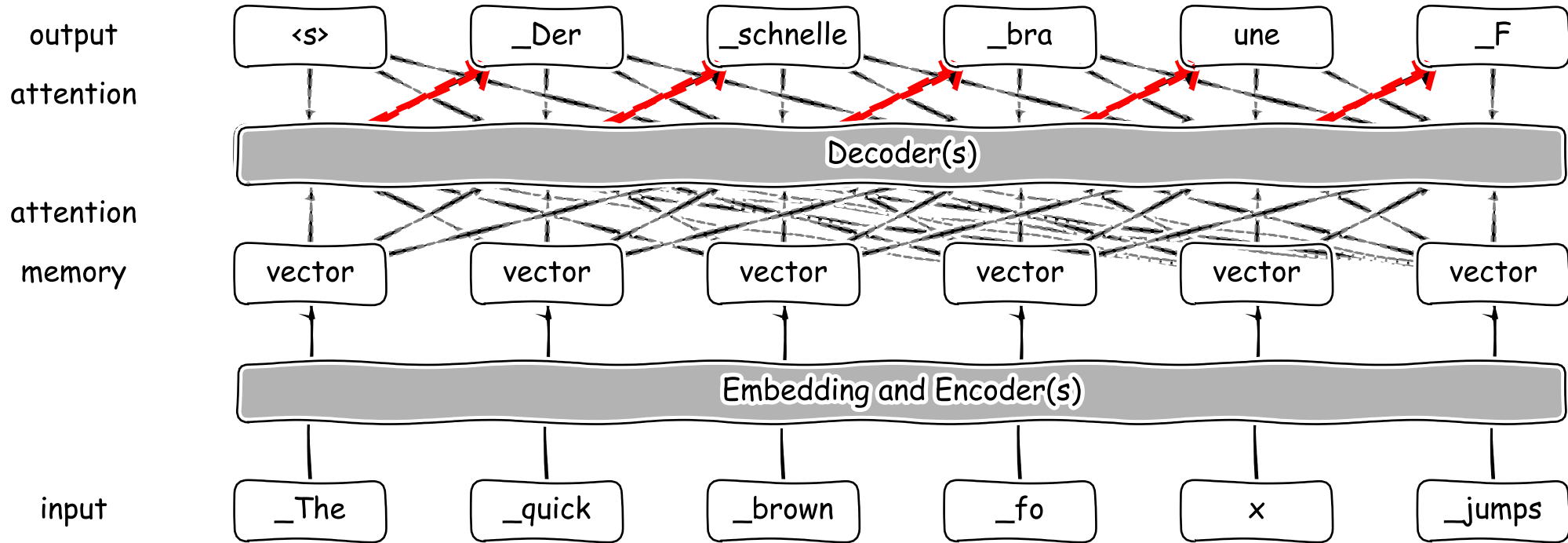


(Source: Strobel et al.: Seq2Seq-Vis)

... but are they useful here without intervention options?



Oversimplified structure of NMT network



- Red arrows: Predictions. Attention "causal" and also in Encoder.
- Similar to the ones used by well-known online services, to recent news-making models (BERT, OpenAI GPT2,...)
- Trained through back-propagation of error term.



NMT models meet the real world

- Need to translate more than one sentence → need to split
- Want to preserve formatting

→ Use Attention

to map

(heuristics

for "end-bias")

→ Paragraph

formatting?

(currently clumsy

bit-by-bit insertion)

	_The	_quick	_brown	_fo	x	_jump	s	_over	_the	_la	zy	_dog	.
_Der	0.07	0.42	0.13	0.03	0.01	0.03	0.02	0.02	0.01	0.02	0.01	0.00	0.24
_schnelle	0.02	0.76	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01
_bra	0.00	0.03	0.90	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
une	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.94
_F	0.00	0.00	0.06	0.02	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
uch	0.00	0.00	0.00	0.75	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04
s	0.01	0.01	0.01	0.02	0.05	0.02	0.01	0.01	0.00	0.00	0.01	0.01	0.85
_spring	0.00	0.00	0.00	0.00	0.00	0.86	0.08	0.03	0.00	0.00	0.00	0.00	0.02
t	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.93
_über	0.00	0.03	0.02	0.00	0.00	0.04	0.03	0.03	0.04	0.11	0.02	0.01	0.67
_den	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.32	0.19	0.01	0.40
_fa	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.08	0.84	0.00	0.03
ul	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.06	0.73	0.00	0.20
en	0.00	0.01	0.03	0.01	0.00	0.06	0.01	0.03	0.01	0.04	0.03	0.06	0.71
_Hund	0.00	0.01	0.01	0.00	0.02	0.01	0.00	0.02	0.01	0.02	0.08	0.04	0.77
.	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.01	0.00	0.01	0.93
</s>	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.95



Building on PyTorch & OpenNMT

PyTorch

- great Python Deep Learning framework,
- quite well-liked in deep learning research,
- large and very nice community

(I'm biased, though.)



OpenNMT

- Quite comprehensive framework for Neural Machine Translation and related tasks
- More library / application like than many other "code for the paper" style implementation.





Training new models

- Needs parallel corpus of sentences (no word alignment needed)
EN-DE: 4.5 Million pairs
- Needs (at least one) "gaming" GPU – for 1-2 weeks
(that is 30-70 KWh per model – compare to ~2.400 KWh per year for a family of 5 - and the GPU)
- Vocabulary preparation
- Training (this is what takes long)
- Evaluation (mostly "closeness" on a holdout set + inspection)
- Probably also want domain adaptation (i.e. specialize from "general" model to one specific for a domain, e.g. legal text).



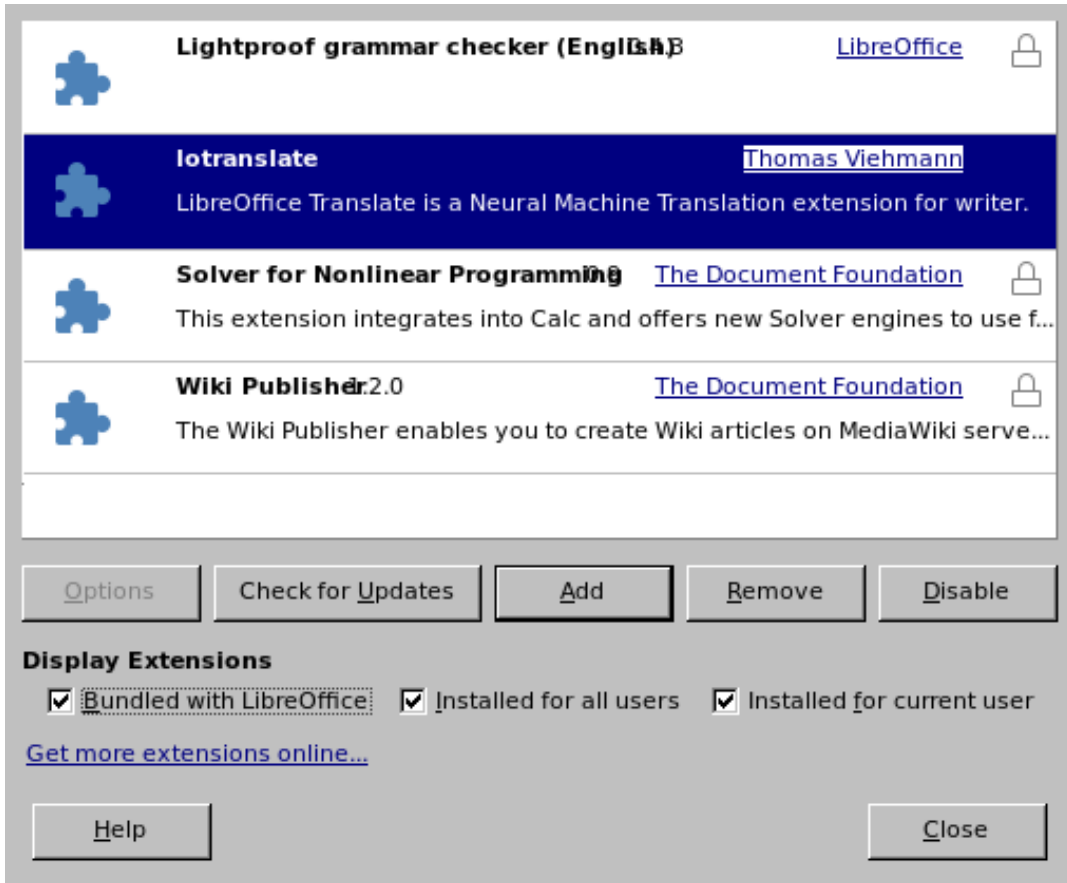
Packaging the extension

- Quite a few Python modules as dependencies: OpenNMT, PyTorch, SynTok (to get sentences) + quite a few indirect ones
- Some are platform / Python-Version specific
- Current solution: build "simple" OXT
- Install dependencies in fresh LibreOffice with pip
- Copy site-packages into OXT zipfile (30MB → 150MB)
- Automate OXT build?

Probably want to shed some dependencies, maybe port to C++.



UI - Installation



- Standard Extension
- (Only) for Windows:
all dependencies bundled



UI – Installation of Translation Models

Install translation models (unpacked from ZIP file) in the in a new Options page

LibreOffice

- User Data
- General
- View
- Print
- Paths
- Fonts
- Security
- Personalization
- Application Colors
- Accessibility
- Advanced
- OpenCL
- Load/Save
- Language Settings
 - Languages
 - Writing Aids
 - Translation**
- LibreOffice Base
- Charts
- Internet

Add source language annotations

Translation models:

- German to English

New Edit Delete

Help Reset Apply Cancel OK



Better integration with LibreOffice?

Some things I would like your input on:

- How much is MT a desired standard feature?
- Working on other than Writer text?
- Do we need more UI?
- Interest in more languages?
- What size of language models is acceptable?
(x.xGB per language pair) Distribution channel?



How to make this more useful?

- Better integration with LibreOffice Writer (e.g. set the Language properties...)
- Would you train a model for your language (pairs)?
What do you look for for training models?
(except a speedup / being more economical)
- Open Corpora: <http://opus.nlpl.eu/>
(incidentally uses translated UI strings as a source)
- NMT researchy (or catching up with NMT research):
 - Can we use dictionaries, too?
 - Improve using “weakly aligned” texts? (from gutenber.org, LibreOffice documentation?)
Related research: automatic filtering noisy corpora (WMT challenge 18 & 19)

Your other comments and observations

Thank you!

Contact:

Thomas Viehmann

<https://mathinf.eu/>

<https://lernapparat.de/>



All text and image content in this document is licensed under the Creative Commons Attribution-Share Alike 4.0 License (unless otherwise specified). "LibreOffice" and "The Document Foundation" are registered trademarks. Their respective logos and icons are subject to international copyright laws. The use of these thereof is subject to trademark policy.