Data Context for Machine Translation

– Conceptual Overview & Resources

1 Introduction

In this conceptual overview, we provide some general information on machine translation-specific data literacy and refer you to some relevant resources outside of DataLit^{MT}, thus establishing a high-level *data context for machine translation* (MT). Data literacy as a transversal skillset for citizens of modern datafied societies has already been explored in depth by various academic and commercial organisations and individual researchers. The same goes for specific aspects of machine translation (literacy), which DataLit^{MT} relates to specific aspects of data literacy (such as ethical aspects of MT). DataLit^{MT} does not intend to reinvent the wheel. Therefore, we refer you to these external resources on data literacy and MT (literacy) in this conceptual overview. We would like to stress that these resources are relevant to a comprehensive (MT-specific) data literacy education, to which DataLit^{MT} only contributes specific building blocks.

2 Basic Information on Data Literacy

The *Arizona State University* provides a 10-minute introductory video on data and data literacy, which provides a good overview of the topic. Arizona State also provides a more comprehensive Study Hall Data Literacy, which you may also be interested in.

You can also have a look at *Jordan Morrow's* 13-minute TED talk on "why everyone should be data literate". In this talk, Jordan stresses the high relevance of data literacy at an overall societal level.

TH Köln, where DataLit^{MT} is based, also hosts a more comprehensive Data Literacy Initiative (DaLI), which develops a full range of data literacy resources not specifically tied to machine translation. You are welcome to have a look at the DaLI resources as well, however, they have been developed specifically for German audiences. For example, DaLI provides a basic course on Data Literacy hosted on *AI Campus*. This online self-study course covers theoretical and practical aspects of the Data Literacy Lifecycle, including establishing a data culture, processing and preparing data, managing data, evaluating and interpreting data, data-driven decision making, data preservation, publicising data and data re-use. The course comprises eight 2-hour sessions and requires German-language skills to complete.

3 Introductions to Natural Language Processing (and Artificial Intelligence)

AI Campus offers numerous online self-study courses and videos – many in German but a selected few also in English – on different topics related to natural language processing (NLP) and artificial intelligence (AI). For example, the *German Research Center for Artificial Intelligence* and *Technische Universität Berlin* created an English-language course on Natural Language Processing. This course introduces NLP using machine learning (ML) and deep learning (DL) models, also covering topics like text pre-processing, vectorisation, text classification and summarisation.

Stanford University provides a comprehensive advanced course on NLP with Deep Learning. The course comprises 23 lectures (each around 1.5 hours), which are all freely available on YouTube.

DeepLearning.AI also provides a range of free courses that may be of interest to you, both on Natural Language Processing Specialization and on related topics, such as artificial intelligence (of particular interest to you may be introductory course AI for Everyone).

You can also have a look at the EasyAI project by *Claudio Fantinuoli*. EasyAI provides an introduction to AI with a focus on NLP. The EasyAI learning resources were specifically designed for readers from the humanities and are therefore very accessible for non-technical audiences.

4 Resources on Machine Translation

The MultiTraiNMT project focuses on "machine translation training for multilingual citizens" and aims to develop both technical and ethical skills in users of modern neural machine translation (NMT) systems. The project is aimed at students and teachers in the fields of language learning and translation. This 2-page paper gives a quick overview of the project. Part of MultiTraiNMT is the MutNMT platform, which allows users to train and evaluate their own NMT models in an accessible graphical user interface that does not require any programming skills. MultiTraiNMT also includes the open-access book Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence. We highly recommend this book to anyone looking for a comprehensive introduction to modern MT technology. The book does not only cover the technical basics of MT in an accessible manner but also investigates other topics such as ethical aspects of MT, which are also relevant in a wider data literacy context.

4.1 Open-Source NMT Toolkits and Related Resources

There is a range of open-source NMT toolkits for training your own NMT models, for example, the minimalist Joey NMT toolkit for educational purposes and the more advanced OpenNMT toolkit. Using these toolkits requires a certain level of programming skills but few adaptations are required for basic NMT model training (unless you actively choose to change individual model parameters). In our DataLit^{MT} learning resource on NMT Training, we implement NMT model training with OpenNMT in a Jupyter notebook, which allows you to train your own NMT model without the programming skills that would usually be required to do so.

For more advanced levels, Hugging Face is a valuable AI community where you can find tutorial videos, source code and data for many AI-related topics including machine translation. Equally, many researchers share their source code in GitHub repositories that can be re-used and built upon, for example, the Trax NMT toolkit developed by *Google Brain*.

4.2 MT Training Data and Suitable Data Repositories

Data availability and quality play a key role when looking for MT training data. The data must be available (either as free open-source data or as fee-based commercial data) and the quality of the data must be good (at the very least, it must be bilingual and ideally already be sentence-aligned, but a high linguistic quality of the data is also very important.). As a rule of thumb, the better the quality of the training data is, the better the quality of the trained MT model will be. A comprehensive resource concerned with MT training data but also with training data for AI/ ML applications in general is the TAUS Training Data Guide.

An ideal repository of open-source MT training data is the OPUS corpus collection. On OPUS, you can download a vast collection of web-crawled, translated data for a large range of language combinations and domains (e.g., data from Wikipedia, News Commentary, TED talks, or the

European Parliament). The size of the datasets ranges from a few hundred to around 280 million sentences, and the data has already been pre-processed automatically. Another, fee-based, source for downloading MT training data is the TAUS Data Marketplace. Here, you can buy parallel, pre-processed MT training data for specific language combinations and domains. The TAUS data is more domain-restricted and usually smaller in scale than the OPUS data. Prices for these specialised datasets range between a few hundred and a few thousand euros. Domain-specific datasets such as those available in the TAUS Data Marketplace can be used for *MT domain adaptation*, which is covered in more detail in this article.

4.3 Preparing Data for MT Model Training

Once suitable training data has been collected, this data must usually be prepared in a specific way before it can be used for MT model training. The TAUS Ten-Step Guide to Data Cleaning provides a concise overview of the individual steps required for training data preparation (for AI/ML applications in general but also for MT models in particular). Particularly the first five steps of 1) language identification, 2) tokenization, 3) removing duplicates, 4) heuristic rules 5) and sentence embedding are relevant for MT training data preparation. We will cover these steps in more detail in our learning resources on Data Planning and Collection: MT Training Data Preparation.

4.4 Conferences and Associations

All over the world, there are numerous conferences on the topic of MT (and NLP in general) each year, hosted by different organisations. For example, there is the yearly Conference on Machine Translation (formerly *Workshop on Machine Translation*, *WMT*), which is sort of the 'main event' for MT and MT research. The conference includes competitions (so-called *shared tasks*) on different aspects of MT. If you would like to know more about this important MT conference, have a look at the website of the Seventh Conference on Machine Translation (WMT22). Then, there's the yearly EAMT Conference hosted by the European Association of Machine Translation and the biannual MT Summit (hosted by different organisations), where researchers and institutions present their most recent research in the field of MT. Further important organisations in the field of MT and NLP in general are Association for Machine Translation in the Americas (AMTA), the Asian & Pacific Association of Machine Translation (AAMT) and the Association for Computational Linguistics (ACL). The ACL hosts the ACL Anthology, a comprehensive archive of papers from the fields of MT and NLP.