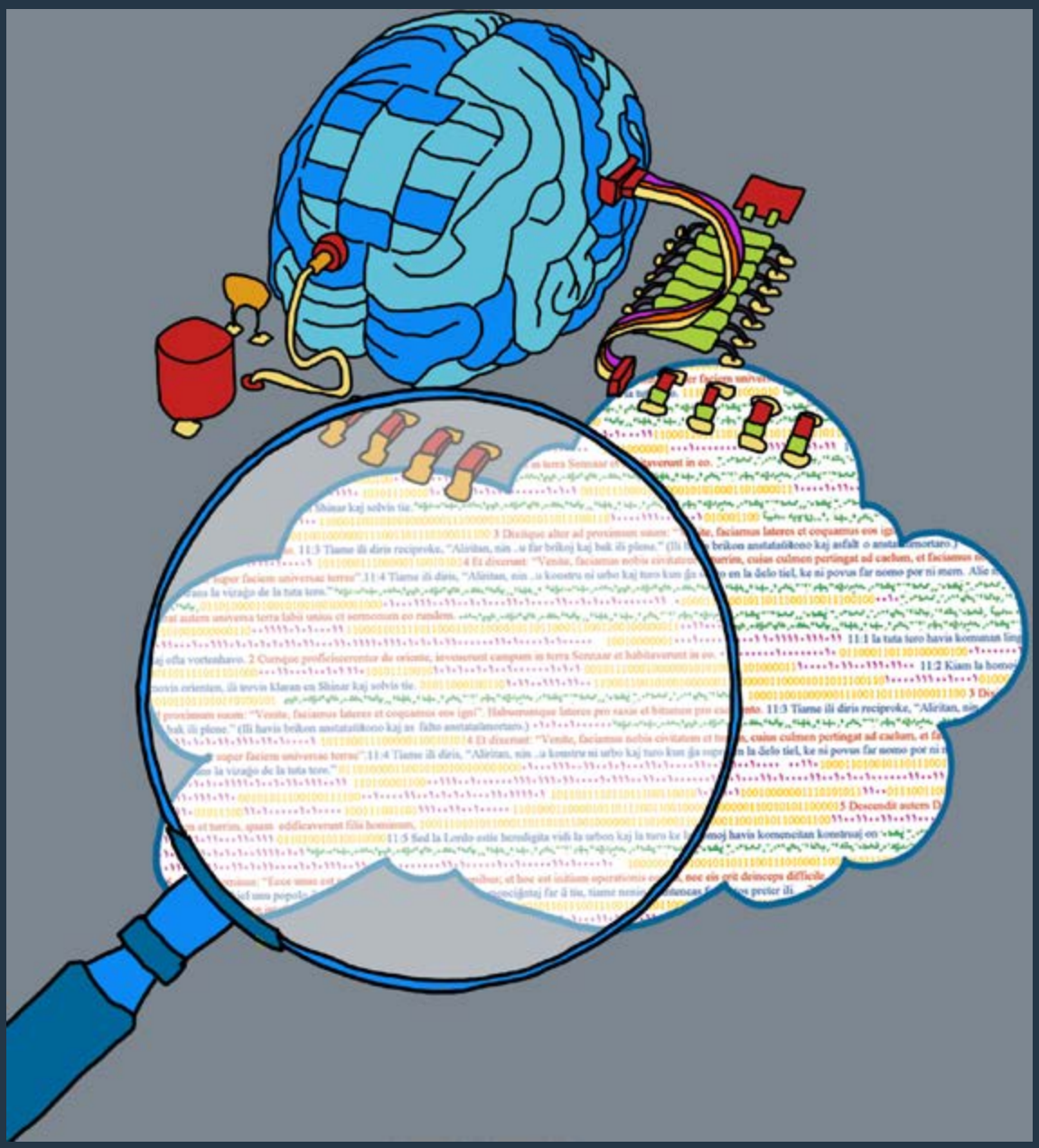
TAUS Matching Data

A new technique to optimize data selection for machine translation training

A TAUS White Paper



January, 2019

TAUS
the language data network

Design: Anne-Maj van der Meer

Published by TAUS BV, De Rijp, The Netherlands

For further information, please email data@taus.net

Table of Contents

Abstract

This white paper describes the challenges of selecting data for the training of Machine Translation systems, especially in light of the transition from Statistical MT to Neural MT. The Neural MT engines demonstrate dramatic improvements in quality of output, but they are much more sensitive to the quality and relevance of the training data. This document gives a complete overview of platforms and organizations that offer data for MT training and describes the constraints that MT users and developers are confronted with. It then presents a new technique developed and implemented by TAUS, based on a joint research project with the Institute for Logic, Language and Computation of the University of Amsterdam, which we refer to as Matching Data. The Matching Data method inverts the typical search approach by indexing all the sentences in the mixed domain search corpora as searchable 'entities'.The resulting domain-specific corpora are more compact and relevant, as demonstrated by the first experiments conducted in close cooperation with the Oracle International Product Solutions group.

The Problem with Language Data

The rapid adoption of artificial intelligence and machine learning in a wide range of products such as automobiles, chat bots, home and wearable devices and a plethora of customer, citizen and patient interaction applications drives the need for language data. The algorithms may be smart, but they only become really useful once they are trained on the languages that users speak everyday.

The primary source for language data has been the world wide web. The Common Crawl Foundation has helped enormously to make massive text corpora in many languages accessible to developers of machine translation. However, the new generation of Neural MT engines are more sensitive to the quality of the language data and they tend to learn better from more domain-specific data. Existing methods of using crawled data are like a scattergun approach to solving the problem.

Whenever possible MT developers fall back on using translation memories for the training and customization of their engines. The problem there persists that the example data may be distributed over a variety of domains and therefore generate very mixed results. The old paradigm of more data is always better seems to start tumbling in this new age of AI. More domain-specific data is definitely better, but how do we tune in to the most relevant data?

More and more developers turn to crowdsourcing solutions for the collection of language data. They provide a list of source segments and pay native speakers and translators for typing in (or dictating) the translations. This method of collecting useful language data seems to be effective, but it's expensive and it may be slow. Also, while zooming in on the exact required vocabulary of an application, this method may be too restrictive for a real-world environment where variances form the rule rather than the exception.

To address the problematic area of language data collection TAUS is launching Matching Data, a new service based on a unique high-performance clustered search technique.

History of Language Data

Overview

This section contains a helicopter overview of initiatives, projects and platforms over the years focusing on collecting and making language data available and accessible to the community.

How Data Entered the Field of Translation

Data entered the field of machine translation in the late eighties and early nineties when researchers at IBM's Thomas J. Watson Research Center reported successes with their statistical approach to machine translation.

Until that time machine translation worked more or less the same way as human translators with grammars, dictionaries and transfer rules as the main tools. Syntactic and rule-based Machine Translation (MT) engines appealed to the imagination of linguistically trained translators, while the new purely data-driven MT engines using probabilistic models turned translation technology into an alien threat for many translators. Not only because the quality of the output improved as more data were fed into the engines, but also because they could not reproduce or even conceive of what was really happening inside these machines.

Google researchers, who also adopted the statistical approach to MT, published an article under the prophetic title "The Unreasonable Effectiveness of Data". Sometimes even the statisticians themselves wondered why metrics went up or down, but one thing seemed to be consistently true: the more data, the better the quality of the translation output.

The English-French Google machine translation engine was trained using a corpus of 100 billion words. Now, with the new generation of Neural MT, very large quantities of data belong to the past.

Early Platforms for Language Data

The Linguistic Data Consortium (LDC) was founded in 1992 as an open consortium of universities, libraries, corporations and government research laboratories to address the data shortage that language technology research and development was facing. LDC's role was initially as a repository and distribution point for language resources. Nowadays it creates and distributes a wide array of language resources.

The European Language Resources Association (ELRA) was founded in 1995 as a non-profit organisation with a mission to promote language resources and evaluation protocols for the Human Language Technology (HLT) sector. As a language resource center, it offers language data such as text and audio corpora, lexica and terminology for HLT to the community.

More specifically it coordinates and carries out identification, production, validation, distribution, standardisation of language resources, and supports the evaluation of systems, products, and tools related to them. ELRA's distribution agency is ELDA (Evaluations and Language resources Distribution Agency), a company that deals with the commercial and business-oriented tasks of the association.

The Open Language Archives Community (OLAC), was founded in 2000 as an international partnership of institutions and individuals that are creating a worldwide virtual library of language resources by developing consensus on best current practice for the digital archiving of language resources as well as a network of interoperating repositories and services for hosting and accessing such resources.

TAUS Data Cloud

In 2008, TAUS and its members founded the TAUS Data Association: a cloud-based repository where everyone can upload their translation memories and earn credits to download data from other users. The objective of the data sharing platform was to give more companies and organizations access to good quality translation data needed to train and improve their MT engines. As of September 2016, the TAUS Data Cloud, as it is called now, already contains more than 35 billion words in 600 language pairs. The hunger for data seems to be unstoppable.

Connecting Europe Facility

In 2014, the European Commission launched an ambitious program under the name Connecting Europe Facility Automatic Translation (CEF.AT) that aims, among others, to promote and support the collection of translation data for automated translation. EU institutions and public administrations in the Member States will have access to an Automated Translation core platform, built on the existing European Commission Machine Translation service (MT@EC), that will automatically translate texts into any EU language.

Member States will be encouraged to contribute language data for improving the translation quality produced by the CEF.AT platform in their language. Within this framework the EU launched a European Language Resource Coordination effort to identify and collect language data from national public services, public administrations and governmental institutions across the 30 European countries participating in the CEF programme. In a parallel move, the Language Resources Observatory site was created to centralize information about "validated" language data resources likely to support Europe-wide automated translation.

Web Crawling

Data is harvested in different ways. Besides the sharing platforms, such as TAUS Data Cloud and CEF.AT, large companies with the proper tools and means have been scraping data from translated websites. Practically all publicly available data resources have been obtained by targeted crawling of a handful of web sites, such as the European Parliament, the United Nations, and volunteer translation efforts such as TED talks and Open Subtitles.

In contrast, the highest-resourced machine translation efforts (Google and Microsoft) use significantly more data by utilizing broad crawls of the web, targeting millions of web sites. By participating in the ModernMT and CEF projects TAUS aims to make such data available for academic research, and governmental and commercial deployment of machine translation with a focus on European languages and Russian.

Embedded Sharing

As in many other private and business environments, users inevitably share data, whether consciously or unconsciously, when they use online services. In the translation industry this happens for instance when translators post-edit MT output on a cloud-based translation platform.

The company that owns the platform or the MT developer that provides the API to facilitate the MT service will receive the translation data and may use this to further improve the performance of their MT engines. These practices raise concerns around copyright issues.

Public Data Collections and Academic Platforms

Public data sets, either standalone or included on academic platforms, that are released by international organizations and/or extracted via academic initiatives, have proven very useful for the advance of machine translation technology, both in research and development. Some of them are described here:

The Europarl corpus is a parallel corpus created by Philipp Koehn from the European Parliament proceedings in the official EU languages. The aim of the extraction and processing was to generate aligned text for use in statistical machine translation systems.

The DGT-Translation Memory owns voluminous data sets released by the European Commission. They contain sentences translated by human translators in 24 languages in the legal domain from the Acquis Communautaire, the body of European legislation. The United Nations parallel corpus is comprised of publicly available official records and other parliamentary documents of the UN in their six official languages. The purpose of the data sets is to provide access to multilingual language data and facilitate research and development in machine translation and other natural language processing tasks.

The Canadian Hansard corpus consists of parallel texts in English and Canadian French, extracted from official records of the proceedings of the Canadian Parliament.

Opus is a growing collection of parallel texts taken from the web. The Opus project aims to convert and align freely available online data, adding linguistic annotation and making available the parallel datasets. It contains data from the European Commission, the European Parliament, European Medicines Agency, the United Nations, Open Subtitles, TED Talks, Wikipedia and others.

Other Translation Data Repository Platforms

Besides the industry-shared TAUS Data Cloud, a number of other translation data repository platforms are available to support and facilitate translation tasks:

- Translated's MyMemory, the world's largest collaborative translation archive, supporting translators and translation automation tasks.
- Linguee, a unique translation tool that combines a bilingual editorial dictionary and a search engine allowing to search for words and phrases in hundreds of millions of parallel texts mainly sourced from public content (EU, UN, etc.).
- Glosbe, a free multilingual translation memory engine populated with open source content (wiktionary, open subtitles, etc.).
- Reverso in Context, a multilingual natural language search engine applied on parallel big data that offers translations in context with real-life examples for millions of words and phrases.
- Tatoeba, a collaborative, open and free collection of sentences and translations.

New Language Data Companies and Services

New companies have emerged that specialize completely in data harvesting. Appen, headquartered in Australia, is the most prominent provider in this niche. But also existing players in the localization space establish new practices directed to data collection. Lionbridge's data business for instance competes with Appen. And there are many more smaller data providers with a localization or crowdsourcing background, such as GlobalMe and Flitto.

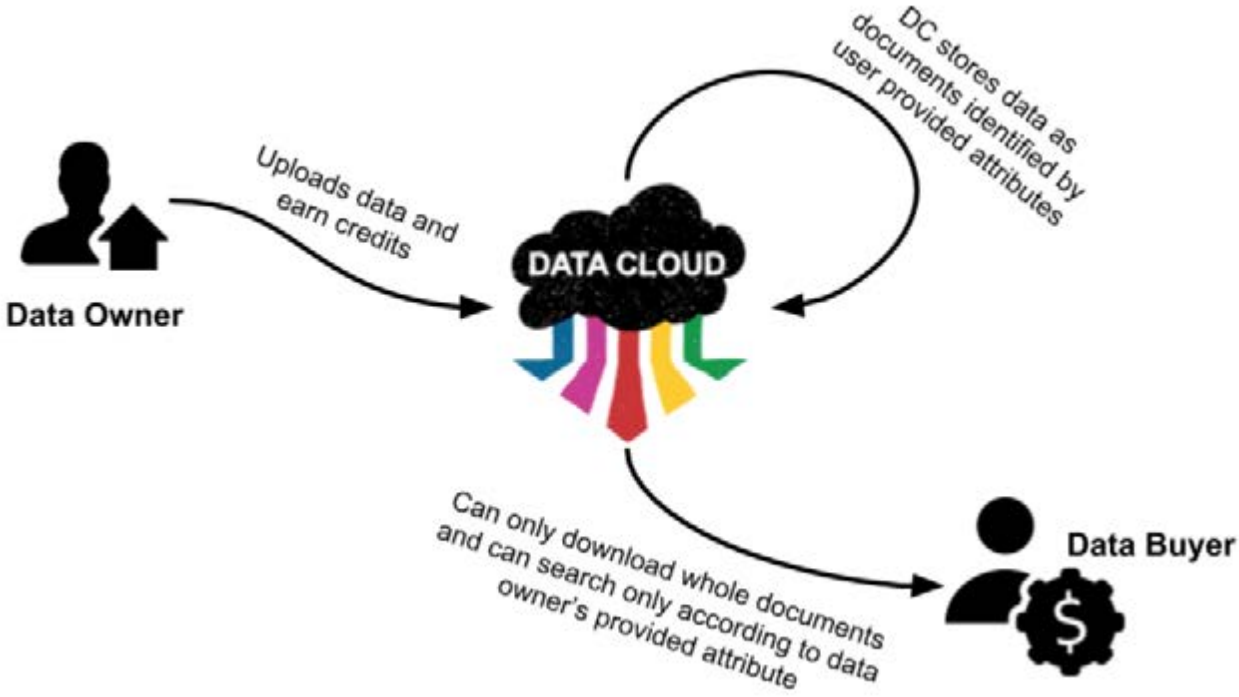
A number of companies in China have also been active in developing large language data repositories, a fact that indicates how vitally useful language data are for the translation industry and other natural language processing applications.

Shanghai-based UTH International has invested deeply over the past few years in building and compiling a large multilingual big-data platform that includes several billion translation units in different domains and using them for a number of multilingual information processing applications and solutions.

Another Shanghai-based company, Tmxmall, has developed a cloud-based platform for translation memory sharing, exchange, trading and other data-related features, following the TAUS Data Cloud example. The goal is to develop a marketplace where buyers and sellers can trade their translation memories to cut translation costs, save time and improve quality.
Baidu, the most popular search engine in China, is, according to Slator, intensively collecting high quality human-translated translation data for translation or other purposes.

Constraints on Language Data Collection

Despite the long list of language data initiatives, platforms and services as outlined here above, access to language data remains problematic. According to a survey that TAUS conducted in 2015 among many data users and producers the primary concern they have is how to get access to exactly the right data. Searching, finding and sampling language data are not simple processes on all of the available platforms, they say. There is a growing need, even more so today with the Neural MT engines, for domain-specific and even customer or application-specific language data. The available services return often collections of data that are too large and too scattered, and as a result these datasets also then come at a high cost compared to benefits. Establishing the right price for language data is experienced as another challenge, especially since the expected improvements in performance of the MT engines cannot be measured until after the fact.

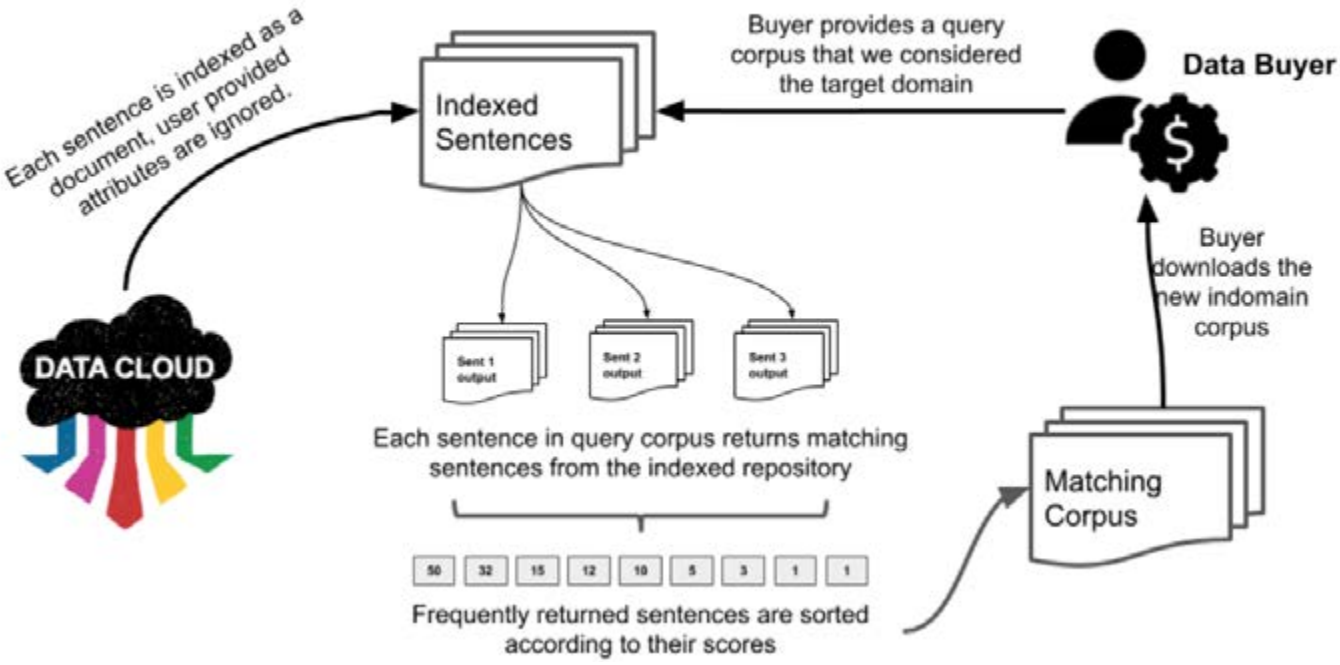


The reciprocal model of the TAUS Data Cloud has served users of TAUS very well in the last ten years. Whoever is in need of data is invited to upload some of his own data. The credits earned as a result of the data uploads can then be used to download other users' data. The constraint that the users experience though is that selections of data can only be made based on attributes that are chosen by the data uploaders from a pre-configured drop-down list provided in the user interface. There is for instance a list of seventeen industry sectors to choose from. That makes it impossible to really tune in on the needs to cover a very specific domain. Another constraint in the old TAUS Data Cloud, as in all other language data repositories, is that users can only download complete files and datasets in exactly the same size as they were uploaded by the data providers. The attribute that was provided for the dataset is automatically associated to each segment in that dataset. As a consequence TAUS Data customers until recently missed opportunities to find and harvest good quality data hidden in datasets with different attributes.

Solution

TAUS is launching Matching Data as a solution to the constraints described above. Matching Data is a high-performance clustered search methodology, based on data selection techniques developed in the DatAptor project.

Here is how it works: based on a sample mini-corpus, the Query Corpus, we identify the best matching data, on a segment-level basis, from the entire data repository. To make this new search and selection method work, each segment in the TAUS Data Cloud is indexed as a searchable document. Similarly, each segment in the Query Corpus is converted into indexed sentences. Each sentence in the Query Corpus then returns matching sentences from the indexed TAUS Data Cloud repository, sorted according to their matching scores. We can increase or decrease the selections by adjusting these matching scores.



The customized corpora built with the Matching Data service are collections of segments extracted from all datasets in the TAUS Data Cloud. This tailored approach to data selection significantly reduces the data volume requirements and finetunes the MT training process. By way of reference: the sample mini-corpus, provided by the userm should contain around twenty thousand segments, monolingual or bilingual, representative for the specific domain. Experience so far has shown that we can condense the data selection by a factor of ten.

The TAUS Matching Data service is available as of January 2019. Users of the service are invited to submit a Query Corpus with a short profile description of the domain. TAUS will then initiate the clustered search process and share the Matching Data corpora in a library on the TAUS Data Cloud. Users can view samples and choose to download compact, medium or large selections of the Matching Data in their domain.

Conclusion

In October and November of 2018 TAUS has tested the new Matching Data service in close cooperation with Oracle International Product Solutions. The process consisted in Oracle IPS supplying TAUS with a sample of approximately 30,000 English strings, representing content that is aligned to Oracle projects.

TAUS used the sample, the Query Corpus, to explore Data Cloud for similarity and proximity, across five languages, and reverted back with three selections of data output, with score ranges on similarity and proximity. Oracle IPS then performed a linguistic assessment of this output.

The in-depth linguistic review rendered positive results and the content supplied by TAUS was of good quality, appropriate to consume as aligned corpora to that supplied in the Oracle sample with an average acceptance score of 84% across the five languages.

TAUS will roll out Matching Data as a new managed service on the TAUS Data Cloud from January 2019. Later in the year, it will be available as a fully automated service allowing users to build domain-specific corpora on demand. Automated cleaning will be applied before the corpora are offered to all users in the Matching Data Library. TAUS plans to extend the clustered search to crawled data from the web. Later in 2019 the Matching Data service will be enriched with an anonymization service and a Data Test Bed, allowing users to validate the created corpus with Matching Data service on their own test sets with NMT models trained on the Matching Data corpus.

See article Fixing the Data Gap for an overview of the Matching Data Roadmap.

References

DatAptor project: https://slpl.science.uva.nl/projects/dataptor/overview

The Unreasonable Effectiveness of Data (2009): https://research.googleblog.com/2009/03/unreasonable-effectiveness-of-data.html

The Human Language Project: Inventing the Future of Translation Data (2012): https://www.taus.net/think-tank/articles/translate-articles/the-human-language-project-inventing-the-future-of-translation-data

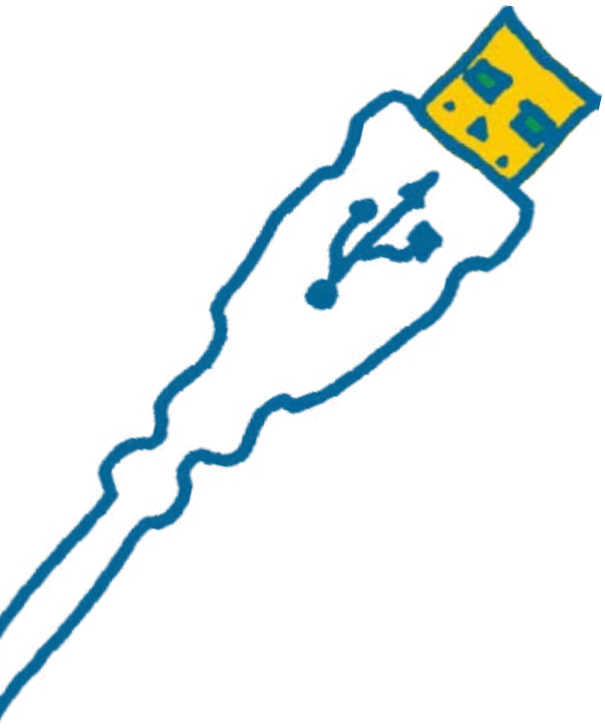
The Call for the Human Language Project (2012): https://www.taus.net/think-tank/articles/translate-articles/the-call-for-the-human-language-project

It's Time for a Big Idea: the Human Language Project (2013): https://www.taus.net/think-tank/articles/event-articles/it-s-time-for-a-big-idea-the-human-language-project

TAUS Translation Data Landscape Report (2015): https://www.taus.net/think-tank/reports/translate-reports/taus-translation-data-landscape-report

TAUS Machine Translation Market Report 2017: https://www.taus.net/think-tank/reports/translate-reports/taus-machine-translation-market-report-2017

TAUS
the language data network



TAUS, the language data network, is an independent and neutral industry organization. We develop communities through a program of events and online user groups and by sharing knowledge, metrics and data that help all stakeholders in the translation industry develop a better service. We provide data services to buyers and providers of language and translation services.

The shared knowledge and data help TAUS members decide on effective localization strategies. The metrics support more efficient processes and the normalization of quality evaluation. The data lead to improved translation automation.

TAUS develops APIs that give members access to services like DQF, the Quality Dashboard and the TAUS Data Market through their own translation platforms and tools. TAUS metrics and data are already built in to most of the major translation technologies.

For more information about TAUS, please visit: https://www.taus.net