RECOMMENDATIONS CONCERNING MACHINE TRANSLATION AND LANGUAGE RESOURCES FOR MACHINE TRANSLATION

Contents

This part of the EcoGuide provides guidance and recommendations concerning machine translation (MT), and language resources (LRs) for MT. For each topic there is a question, an answer, and practical considerations, organized according to the needs and interests of different target groups.

This part is recommended for:

- End users
- LTs developers/providers
- LRs developers/providers
- Policy/decision makers

Q1: How to use translation technologies?

A: Translation technologies have been primarily developed to translate written data, i.e. turning a text in one language into a text in another or several other languages. Translation technologies broadly comprise two different types of technologies:

- Machine translation (MT), often called automated translation, aiming at automatically or semiautomatically translating text or speech, with several different approaches:
  - (depending on the degree of 'automated'): Fully automated machine translation, humanassisted machine translation (e.g. post-editing)
  - (depending on the technological approach): Statistical MT, rule-based MT, neural MT, adaptive MT, hybrid MT systems;
- Computer-assisted translation (CAT) with various CAT tools: CAT tools often comprise translation memory modules, which are a sort of LRs, and terminology database modules.

Increasingly, translation technologies are also applied for:

- Interpreting spoken data into different languages
- 'Transcreation', i.e. the translation not only into one or more different languages, but also into different sorts of texts (i.e. for instance reformulating a text addressing a specialist audience into a text addressing a non-specialist audience)

MT systems can be combined and integrated with:

- Corpus technologies using mono-, bi- or multilingual corpora for supporting MT or training MT systems,
- Tools for mono, bi- or multilingual LRs containing structured content needed for MT or directly to support MT,
- CAT tools,
- Authoring tools, technical documentation systems, etc.

Integration of these diverse tools remains a challenge even for mature adopters of MT. Text generated with these tools may include formatting codes and tags, which are reportedly difficult to deal with in an MT application and need to be often engineered around.

Adding to the complexity, a wide range of tools may be needed, such as aligners, taggers, tokenizers, lemmatizers, chunkers, parsers, disambiguation tools, concordancers, annotation tools, extraction tools, etc. for pre-processing and cleaning the data (e.g. when acquiring in-domain parallel data from the web, but also customer's data) for various MT-related processes. However, the results of applying this technology may be particularly useful not only for MT, but also for information gathering and business intelligence.

What to consider:

Fitting translation technologies well into the workflows and communication flows of an organization is decisive for the efficiency and effectiveness of the technology applied.

Consider the reuse and repurposing of the data generated with translation technologies, but also other technologies (e.g. word processors, authoring tools, desktop publishing systems, technical documentation systems, etc.) as an element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector data would add an enormous potential of LRs in different field. Opening up these sources would help MT at large.

The efficient use of one or a combination of the translation technology tools/systems depends on several conditions: purpose of the translation, expected quality, language pairs, domains, text types, time and cost factor, volumes to be translated, re-use of existing LRs in the organization, need for brand fidelity, privacy and data security issues, role of translation technologies in the information flows of the organization, etc.

Thoroughly check integratability and interoperability aspects, especially 'content interoperability', if content is planned to be used or reused in more than one translation technology application, in other technologies, or in value-chains. Standards can greatly support implementing interoperability.

Q2: Do tools/systems need to be interoperable?

A: State of the art language technologies (LTs) should be technically interoperable, as they are increasingly needed in some or the other combined form. Besides, as quite some kinds of language resources (LRs) can be used across some or many of the above LTs, interoperability is a serious issue.

What to consider:

Be sceptical and scrutinize statements, such as "this LTs tool/system is fully multilingual", "this LR can be easily converted into other languages", "this LTs tool/system is interoperable with other tools/systems", this LR is interoperable with other LRs". "99% interoperability" often means in fact noninteroperability!

Interoperability should be achieved along three layers of interoperability: technical, organizational, and semantic interoperability. Standards can greatly support implementing interoperability along these three layers.

Organisations listed in European Directory of Language Technology Vendors and LT-Observe directory may assist in putting the right questions.

Q3: Does MT provide a good enough translation quality?

A: A better question would be "a translation good enough for what?"

Depending on the intended use of the output, the translation quality may be satisfactory for:

- somehow understanding a text? – gist translation, for instance real-time communication, such as online chats
- certain sorts of texts only? – for instance technical manuals already written in highly controlled language
- producing 'publishable quality' texts? – for instance for product catalogues produced by extracting data from product master data management systems, or texts that must meet special requirements (e.g. legal requirements concerning liability)
- translating texts without post editing? – depends on which kinds of text and for which purpose as well as for whom

Not all content, not all file formats and not all language pairs are suited to the same engine: The expected purpose of translation in relation to the required quality, content types and language pairs tend to play a determining role in engine performance. Satisfactory results can be achieved in certain domains (e.g. technical, administrative), and text types (e.g. patents, technical manuals, software documentation, vehicle assembly build instructions). Rule-based MT systems reportedly perform better in so-called "narrow domains" and certain language pairs (e.g. Japanese-German), while

statistical MT systems are better suited for "broad domains" (e.g. user generated content) and languages of lesser distribution. Regardless the "tool wars", often an adequately performing MT system that fits into the workflow is preferable in terms of costs, time factor and benefits to a better performing MT that is not integrated into the workflow.

What to consider:

Determine the volumes of text, content types and language pairs to be machine translated for which purposes in relation to the expected or required quality of outcome – and the expected benefits under the consideration of time and cost factor.

Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

Q4: How to apply machine translation (MT) optimally in the organization?

A: The best machine translation solutions raise output quality, lower the human factor across the spectrum of applications, and therefore improve cost-effectiveness and availability of translation services. MT may be also capable of delivering brand fidelity through implementation of organizationspecific language, while at the same time reducing the need for human post-editing, by aligning translations with organization's corporate language.

Many people use MT freely offered through the Internet for something like 'informative raw translation' (gist translation). This is different, if MT is used as one of the central systems in the organization: security and confidentiality, and reliable use of client's data and corporate terminology are the main benefits of customised MT service in professional environment.

What to consider:

To implement MT effectively in professional environments, it is important to see MT as a process. This requires a thorough analysis of how to integrate MT optimally into your organization's system environment, workflows, etc.

Determining what engine and what process will give the best results, under the consideration of time frame and cost factor and based on the following:

- The expected purpose of translation in relation to the required quality, content types and language pairs, etc.
- Would it need a combination and integration of MT systems with other technologies?
- What are the costs for:
  - the preparation/pre-editing of texts (and which kinds of text?) to undergo MT?
  - the human post-editing process (towards which quality level)?

Language resources (LRs) needed at what costs:

- What kind language resources are necessary to deploy the MT system you would need? Are the LRs available in your organization?
- What are the costs for the maintenance of
  - the MT system (possibly in combination with other language technology tools)?
  - the systems for processing and maintaining LRs?
  - different kinds of LRs for different roles in the MT process?

The integration or outsourcing of MT:

Do you want an internal engine that is customised to your needs, or do you plan to outsource all or parts of MT activities, from customising and processing to post-editing and maintaining?

Often, translation is outsourced (with or without MT). This means that valuable translation data does not become part of an organization's asset. Therefore, integrated solutions may be at the beginning more costly but could add value for future translation towards additional languages.

This issue is to be considered under strategic aspects, such as cost-effectiveness, quality management, legal implications, data security and confidentiality, knowledge of the organization.

Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

The usefulness or even need of a organisational language strategy with respect to:

- formulating (or adapting) a organisation-specific language strategy
- taking into account aspects of legal or technical regulations, customer relations, etc.

Q5: What are the most important language resources (LRs) for use in machine translation (MT)?

A: The term language resource (LR) refers to a set of language or speech data and their descriptions in digital form. They are used for building, improving or evaluating natural language (human language) and speech algorithms or technologies. They are increasingly used for machine learning. Furthermore, they are also widely used in the language industry, in language and translation studies, in electronic publishing, for international transactions, etc.

Whichever type of machine translation model is used, the key to creating a good system is currently lots of (quality) LRs. LRs relevant for MT comprise a broad range of different kinds of structured content and corpora. For text-based data (as opposed to multimodal, e.g. audio or video data) they can be broadly differentiated into:

- Text corpora, such as
  - Parallel corpora, i.e. corpora where the same text appears in more than one language (particularly useful for machine translation)
  - Comparable corpora, i.e. corpora with texts of the same or very similar topic in the same or similar sort of text in different languages
  - Monolingual corpora
- Terminologies and similar data (such as domain-specific thesauri, classifications, nomenclatures, taxonomies, etc.)
- Lexicographical data and similar (such as dictionary data, treebanks, grammars, etc.)

Depending on the purpose of the use, other kinds of structured content (pictorial data, directories of proper names of all sorts, dialogue data, etc.) may also be or become necessary.

What to consider:

Analyse carefully your needs for which kinds of LRs and the languages needed. Organizations listed in LT-Observe directory may assist in finding the right solution for your needs.

Set up the criteria for the usability of the LR for the specific project (language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, etc). Consult the LTObserve Catalogue to obtain operationally usable LRs for commercial purposes.

Consider the reuse and repurposing of LRs within the organization as an essential element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector information would add an enormous potential of LRs in different fields. Opening up these sources would help MT at large.

Q6: What corpora are useful for machine translation (MT)?

A: Most corporate buyers believe parallel data LRs (two or more languages side by side) are what is needed to train a statistical machine translation (SMT) system. However, monolingual data in the target language are becoming increasingly useful. In statistical MT a crucial step is to develop a language model for the target language that selects the best translation. Therefore an MT engine needs to be trained on monolingual data as well as parallel data to produce a fluent output.

What to consider:

Analyse carefully your needs for LRs. Organizations listed in LT-Observe directory may assist in finding the right solution for your needs.

Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, etc). Consult the LT Observe Catalogue to obtain operationally usable LRs for commercial purposes.

MT developers need a considerable amount of data to build an MT system. As a rule of thumb, the closer the data that is used to train the system to the type of data that should be translated, the better the results will be. The universally preferred format is plain text or XML; TMX and XLIFF data formats may be preferable for parallel resources. Reportedly an MT developer may need at least 5 million tokens/500,000 segments of language data to build an MT system for a particular domain. In terms of size and quality, balance is needed.
The widely-used statistical MT, Moses, needs sentence-aligned data for its training process. If data is aligned at the document level, it is recommended to convert it to sentence-aligned data using a sentence aligner. See Best Practice Guide to LRs for Automated MT for a report on two sentence aligner tools, Hunalign and BSA.

More parallel data is needed in all the verticals: legal, tourism, etc., as well as for lesser covered languages. New MT models (e.g. neural MT) may have an impact on the need for data, e.g. for monolingual data.

Q7: How to find language resources (LRs) needed for machine translation (MT) or other purposes?

A: LRs are indispensable for the development of tools for machine translation (MT), but they are also expensive and labour-intensive to create or adapt e.g. for MT usability. It should be noted that the extent of the availability and coverage of LRs differs considerably from language to language.

The larger the organization, the higher the probability that at least some of the LRs needed exist already. However, if existing data are fragmented, partially outdated, mixed with heterogeneous data, or come with legal and confidentiality problems, etc. obtaining them from external sources may be more cost effective and less time-consuming.

What to consider:

Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, etc).

Consult the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes.

Consult the repositories and catalogues, such as those provided by ELRA/ELDA, CLARIN, OPUS, METASHARE.

Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector information would add an enormous potential of LRs in different fields. Opening up these sources would help MT at large.

Q8: How to evaluate the usability of language resources (LRs) needed for machine translation (MT)?

A: There is currently no clear answer from industry about any shared method for evaluating the ***usability*** of LRs. There are many aspects of the usability of language resources, and it only makes sense to talk about usability for a specific task. In general, the basic usability of the resource is evaluated by users based on **language (pair)**, **domain** and some **basic file format metadata**.

What to consider:

Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, file format, etc).

Check the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes.

Consult the repositories and catalogues, such as those provided by ELRA/ELDA, CLARIN, OPUS, METASHARE.

Analyse carefully your (immediate and future) needs for LRs. Check to which degree a LR must be interoperable with other LRs or LTs tools/systems and whether it can be scaled-up. This applies to both commercial or to open source. In spite of existing standards (mostly focusing on technical interoperability), the interchange, re-use and repurposing of LRs is not trivial and may require huge efforts, if not carefully planned. Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

Q9: How to add value to existing language resources (LRs) for machine translation (MT) purposes?

A: Many LRs in existing repositories present some challenges in an operational context. Valorisation of LRs through optimization of existing metadata, and sometimes addition of new metadata, will be an invaluable aid to MT developers - as this means that the developers will actually be able to identify and select the resources they need.

What to consider:

Consider adding the following recommended metadata to LRs to be used for MT purposes:

Title, Resource type, Creator, Language(s), Availability, Modality, URL, Domain, Format, Size, Production date, Comment, Description, Tags, Contact person, Format description. For detailed description and examples, see Best Practice Guide to LRs for Automated MT.

Consider following a metadata standard, for example the Dublin Core metadata set for resources; even though some adjustments are recommended, such as specifying the production date of the resource (*Production date*), inclusion of categories that are not part of the metadata set (*Size, Comment, Modality, Availability, Tags*), and omission of certain categories. For details, see Best Practice Guide to LRs for Automated MT.

Collect best practice information: send LRs with initial valorisation to potential LR users, vendors and buyers and let them examine and test resources in relation to their own work context and requirements. A user-feedback system in a collection can help to continuously improve the resources. A mere rating system may not be useful, as such rating may differ depending on purpose and use of the LRs.

Q10: How to create language resources (LRs) needed for machine translation (MT), especially aligned corpora?

A: Performance of statistical machine translation (SMT) systems is depended on how well the training data correlates with the documents that are translated regarding genre, style and in particularly domain-specific data. As these types of data are often not available, a recommended technique to create new LRs is to collect the data, for example domain-relevant training data, by exploiting webcrawling approaches.

What to consider:

Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, file format, etc).

Check the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes. Consult the repositories and catalogues, such as those provided by ELRA/ELDA, CLARIN, OPUS, META-SHARE.

Be advised that the issues of Intellectual Property Rights (IPR) hamper the free use of materials from the web.

See Q11: What are the best practices to lawfully acquire data through web-crawling? for details.

Acquisition of in-domain parallel data can be divided into three phases:

- A focused search for and subsequently ranking of domain relevant websites. The links found at these websites are then regarded as candidate URL seeds with respect to identifying bilingual documents, and the detected candidate documents are evaluated.
- Cleaning up and preparing documents: Removal of duplicates and exclusion of boilerplate elements.
- The next step consists of sentence splitting and tokenization. The final step in the pipeline of making parallel data qualified as training data for an SMT system, is to secure that the sentences extracted are aligned with the highest quality possible.

See the Best Practice Guide to LRs for Automated MT for detailed instructions for each phase.

Useful open resource tools do exist that can help you through the pipeline steps (see the list of recommended tools), but no single place exists where open source software for generating high quality in-domain and sentence aligned corpora, was available at one single website or portal, requiring that the users themselves are left to hard code the software that integrates the applications into one workflow.

Be advised that the internet increasingly contains content that has already been translated automatically, which is polluting the linguistic quality of internet-based content in general. Currently, there seems to be no quick method for deciding whether a given content found on the internet is "human translated" rather than produced by a machine.

Q11: What are the best practices to lawfully acquire data through web-crawling?

A: The issues of Intellectual Property Rights (IPR) hamper the free use of materials from the web. The legal framework within which agile corpus acquisition would operate is governed by two sets of legislative provisions: copyright and database rights (intellectual property rights, IPR), and data protection (privacy and autonomy, i.e. confidentiality, anonymity and access arrangements). In some countries there is copyright exception for specific purposes. Certain materials exist which do not fall under the copyright protection law – this mainly concerns texts from public administration, which can therefore be very useful in this context.

What to consider:

It is recommended using data where the rights are cleared – e.g. LRs from ELRA, and other providers (CLARIN, META-SHARE, etc.) where license is regulated. These materials for the most part come with a price tag and/or licensing conditions governing their use. The existence of license conditions attached to the use of a specific resource means that an agreement between the IPR owner and the provider, e.g. ELRA (see http://wizard.elda.org/principal.php), has been entered into about the conditions regarding distribution and use of the resource.

The resources listed in LT-Observe Catalogue have been included based on the above considerations for commercial purposes.

The process of assessing whether to harvest Web data and obtaining permission to use these can be split up into three steps:

- Locating the data relevant for your MT systems in terms of number sources and especially what means and tools that are needed to handle these data (see above).
- Determining execution costs, i.e. will it, seen from a cost-benefit point of view, be worthwhile to conduct the time-consuming analysis of: conditions and terms for using the data at a Web site, and possibility of identifying responsible content providers
- Evaluate the collected information. Be advised that the negotiations about data usage rights can be cumbersome and time consuming to conduct.

See the Best Practice Guide to LRs for Automated MT for detailed instructions for each step.

Q12: What are terminological data and similar data good for in machine translation (MT) and how to generate them?

A: Terminologies are usually following a language-independent approach, which allows managing data in two or more, sometimes many languages. Terminological data are not only semantically structured, but also contain additional information such as definitions (or explanations, contexts, etc.), references etc. Some large terminology collections are monolingual, but most of them are bi- or multilingual. Others look monolingual (such as some harmonizing quantities and units), but they can easily be turned multilingual. Many terminology collections are of a highly-harmonized nature and therefore contain particularly reliable data.

Similarly, structured LRs are domain-specific thesauri, classifications, nomenclatures, taxonomies, translation memories and the like. They too are mostly multilingual and usually contain highly harmonized data.

What to consider:

Consider the reuse and re-purposing of terminologies within the organization as an essential element in the operation, viz. value chains, of an organization. Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

Terminology collections may contain equivalents that are rated as unlikely by the statistical machine translation (SMT) system models. If such an SMT system is integrated in translation service workflow, it is not possible to ensure high quality terminology (consistency, correctness) in the SMT suggestions. Training data can contain contradicting terminology, corporate specific synonyms or brand terminology. For this reason, effective adaptation of SMT systems can profit from customized client's terminology collections.

Clients require correct and accurate use of specific terminology, often corporate terminology. Clients may provide their own terminology collections, but in projects where the client-supplied terminology collections are not readily available and the use of specific in-house terminology is still required, the terminology firstly needs to be extracted from documents provided by the client.

Term extraction generally involves four steps:

- compilation of a specialized corpus,
- extraction of term candidates (useful tools do exist that can help you to extract term candidates, see the list of recommended tools),
- validation of the term candidates and
- automatic or semi-automatic creation of terminological records.

See the Best Practice Guide to LRs for Automated MT for detailed instructions for each step.

Q13: What other structured LRs could be used?

A: The usefulness of other kinds of bi- or multilingual structured LRs, such as bi- or multilingual directories of all sorts of proper names (many names differ, have aliases, or are differently spelled or pronounced in different languages), certain types of master data (e.g. product properties in enterprise resource management systems or more specifically in product master data management systems), ontologies, etc. is generally not yet fully recognized in the field of the LTs. In particular, master data for instance in trade often need to be multilingual – some of them are even standardized and maintained by maintenance agencies or registration authorities. So far, this source of LRs has not been tapped – probably due to difficulties in accessing the respective LRs.

What to consider:

Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. Organizations listed in LT-Observe directory that may assist in finding the right solution for your needs.

Q14: Are LRs used only for translation purposes?

A: Some kinds of LRs also have or can have other purposes, as resources in authoring tools, for technical documentation, procurement purposes, organizational knowledge management, in archives, etc.

What to consider:

Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. Organizations listed in LT-Observe directory may assist in finding the right solution for your needs.

LRs may be created for specific purposes and within particular frameworks, and this is the information that is crucial to pass on to other users of the resources. Here metadata come into the picture; when an LR is provided with sufficient metadata that thoroughly describe the resource, the users can decide for themselves whether it is likely that this resource can be used in relation to the particular task. See Q9: How to add value to existing language resources (LRs) for machine translation (MT) purposes?

Q15: What organizations would be interested to use or repurpose language resources (LRs) for what purpose?

A: This depends on the purpose of the different kinds of LRs. Language service providers (LSPs) may need to use any of the LRs mentioned e.g. for translation purposes (and the respective language technologies). MT developers or MT service providers may primarily be interested in bi- or multilingual text corpora (or parallel texts or comparable texts), terminological data and lexicographical data. Enterprises and public services usually have several or many (not to mention different kinds of) databases with structured content for different purposes, as well as data generated with text processors, word processors, authoring tools, desktop publishing systems, technical documentation systems, etc.

What to consider:

Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector information would add an enormous potential of LRs in different field. Opening up these sources would help MT at large.