Why Life Sciences Is an Ideal Domain for Machine Translation

TABLE OF CONTENTS

INTRODUCTION

There has always been considerable doubt on whether machine translation is suitable for translating life sciences content. Many translators and LSPs maintain that machine translation is not suitable for this purpose because of the mission-critical nature of content that demands extensive human skills and expert domain knowledge. Mission critical is a term they apply to materials that, if mistranslated, could compromise patient safety. Translation errors in this domain can literally be a matter of life and death.

Thus, this is a domain that has always been much more focused on quality control and has always sought out subject matter expert linguists to do a lot of the critical translation work. Often the translators are doctors, nurses or technicians who used to work with the medical equipment and medical specialty that they are translating for. The extensive use of subject matter expert linguists at various phases in the translation process is more extensive in this domain than most others.

However, Omniscien Technologies contends that **many of the very reasons that are provided as objections against the use of machine translation are actually the points that make the Life Sciences domain ideally suited to expert customized machine translation.** These include:

- High levels of quality control on historical human translations produce high quality translation memories that are ideally suited for machine translation training data.
- Terminology has historically been managed and clearly defined. Once this terminology data is applied correctly to a custom machine translation engine, terminology will be corrected every time.

HIGHLIGHTS

• Technology and approaches to machine translation have evolved in recent years to deliver previously unheard translation output quality and also provide a degree of user control that is required in order to be useful in this complex and quality critical domain.

- When the Clean Data SMT model is leveraged, the focus is around quality rather than quantity of data. This methodology for customizing machine translation requires lower volumes of data than other approaches and delivers very high consistency in terminology and writing style.
- High levels of quality control are still employed whether the initial translation is performed by human or machine. The resulting post edited machine translations will improve the quality of the custom engine more rapidly than other approaches due to the quality and consistency of the final deliverable translation.

Life Sciences domain content can vary greatly and the quality requirements can also vary across different kinds of content such as:

- Informed consent forms (ICFs)
- Instructions for use (IFUs)
- Patient-facing materials
- Product literature
- Research material
- Regulatory documents
- Training manuals
- Technical papers
- User manuals

Historically machine translation has been of lower quality and never reached output quality levels that would make it viable for this domain. However, technology and approaches to machine translation have evolved in recent years to deliver previously unheard translation output quality and also provide a degree of user control that is required in order to be useful in this complex and quality critical domain.

DEFINING THE LIFE SCIENCES DOMAIN

Before we explore the use of machine translation in the Life Sciences domain, it is important to define the Life Sciences domain so as to be totally clear about what kind of content is being discussed.

There are a number of common characteristics of translations in the Life Sciences domain when translated by a human:

- Life sciences content is typically expected to be of very high quality.
- Terminology is very well defined.
- The majority of terms are explicit and non-ambiguous.
- Investment in quality control is greater than most other domains.

For the purposes of this article, we will focus on the types of content that need to be very high quality and is often labeled as "mission-critical, i.e. if mistranslated, could compromise patient safety and the consequences of errors are potentially much higher.

DETERMINING THE BEST MACHINE TRANSLATION APPROACH FOR THE LIFE SCIENCES DOMAIN

Now that we have a clear definition of the Life Sciences domain, let us also define machine translation. Machine translation is not a single technology. There are many different machine translation approaches and solutions available in the market today. All too often, when machine translation is mentioned, people think of older rules based technology or generic machine translation like Google Translate or Microsoft Translator. Even when custom machine translation is mentioned, most think of systems built using Moses open source technology that has been built on top of other non-refined out of domain corpus such as Europarl. This approach is called Dirty Data SMT, where data from multiple sources is mixed with little or no control.

Uploading data to an online service or adding data to an in-house Moses system without the ability, knowledge or skills to manage and refine the data is not user empowerment. In many cases the quality of the resulting custom engine will be lower than that of free generic MT such as Google Translate. In a recent case study, IOLAR built their own custom machine translation using Moses over a six-month period to translate from German to Slovenian in the Technical Engineering domain. Ultimately IOLAR turned to Omniscien Technologies when they could not address issues with word order, terminology consistency, unknown words and incorrect inflected forms in the Moses system.

HIGHLIGHTS

• With good expert managed custom machine translation development, not only will users get projects done faster, but lower unit costs can also allow more content to be translated.

• Building high quality custom machine translation systems requires deep understanding and expertise and the outcomes are quite different when these elements are present.

When quality and user control are high priorities, it is critical that deep expertise is applied to the custom machine translation engine development process. For machine translation to be viable in the Life Sciences domain, users must feel that minimum raw machine translation output standards are achievable and that the effort to post-edit the output would actually be less than doing it via traditional Translate, Edit, Proof (TEP) processes. Most do not realize that machine translation simply replaces the T step of TEP and that edit and proofing are still required for final quality. With good expert managed custom machine translation development, not only will users get projects done faster, but lower unit costs can also allow more content to be translated. Building high quality custom machine translation systems requires deep understanding and expertise and the outcomes are quite different when these elements are present.

SUCCESS WITH CUSTOM MACHINE TRANSLATION ENGINES BUILT WITH THE CLEAN DATA SMT MODEL

Numerous Omniscien Technologies customers have successfully developed very high quality custom machine translation engines using Language Studio™ Enterprise and Language Studio™ PowerTrain. The key to their success is the use of the Clean Data SMT model combined with Advanced Data Manufacturing to customize the machine translation engine with guidance from the customer's subject experts and Language Studio™ Linguists.

The Clean Data SMT model of customizing very high quality machine translation engines was pioneered in 2008 by Omniscien Technologies and has 4 key requirements:

1. When customizing an MT engine, the user must be able to define and refine the writing style, preferred terminology, target audience and purpose of the custom engine.
2. All the data used within the custom MT engine must be able to be refined to match the needs and purpose of each individual engine.
3. High quality in domain data should be the primary data used to build statistical models.
4. Cleaning data requires human cognition and understanding of the data.

These requirements are explained in more detail in the article "Clean Data Statistical Machine Translation".

Omniscien™ TECHNOLOGIES

HIGHLIGHTS

• The Language Studio™ approach empowers experts in localizing for the medical industry and gives them complete control of all data within the custom machine translation engine. Because of this deep subject expert engagement, the risk of using machine translation as part of the translation process is greatly reduced and successful outcomes greatly enhanced.

In the Clean Data SMT approach, data is understood by a human specialist and key decisions are made to refine and adapt the data so as to ensure delivery of the correct terminology and writing style. The objective is to enhance and improve the translation process, both in terms of efficiency and effectiveness. Included in this data are customer related translation memories, product reference materials and terminology lists. This is the same set of materials that would normally be provided to a subject matter expert linguist in a traditional human only project. The full set of data that the custom machine translation engine is built upon in its entirety is analyzed and adapted to be compliant with all the terminology and other project requirements through an extensive set of processes that are executed by Language Studio™ Linguists, but guided by the client's subject matter experts.

Refining data for high quality machine translation is a complex task that requires expertise and experience and Language Studio™ Linguists work directly with our customers to refine the data, removing ambiguity and improving quality. With the Clean Data SMT model, terminology is normalized and when there are multiple possible translations for terminology, the client's experts in localizing for the medical industry and the specific customer are engaged to refine and remove ambiguity. Language Studio™ Linguists work with these experts to extract, refine and apply the terminology.

Where there are gaps in data coverage, Language Studio™ Advanced Data Manufacturing is used to produce candidate data which is then validated by the client for accuracy and quality. Subject experts in localizing for the medical industry that work for the end client are heavily engaged throughout the entire lifecycle of the custom machine translation engine.

Because the data used within the custom machine translation engine is very high quality and has been further refined from its original translation memory form, the quality of terminology and translation is greatly improved. The Language Studio™ approach **empowers experts in localizing for the medical industry and gives them complete control** of all data within the custom machine translation engine. Because of this deep subject expert engagement, the risk of using machine translation as part of the translation process is greatly reduced and successful outcomes greatly enhanced.

One of the common criticisms of machine translation is the unpredictability of the resulting translations. Using the Clean Data SMT approach, issues of unpredictability can be greatly reduced by progressively refining and normalizing terminology and then enforcing consistent language patterns with data manufactured around these critical terms. As the primary data that the custom engine was built using the clients own high quality translation

Omniscien™ TECHNOLOGIES

memories that are in a narrow and granular domain, the resulting translation is very predictable. When translation issues do arise, very small amounts of corrective data or a refinement rule will usually address the issue so that it does not occur again in the future.

Due to the human cognition of experts who know the industry and the customer, the data that a custom engine is built upon is of exceptionally high quality. Thus, the resulting machine translation output is also considerably higher than approaches that are not built on the Clean Data SMT model. In many cases we have seen fewer edits during the quality control process than a first pass human translator.

MANAGING RISK

Because of the quality level required for "mission-critical" translations, a rigorous quality control process is essential whether using human or machine translation. It is important to understand that the machine translation component is only used at the same stage as a first pass human translator.

The materials go through a translate/edit/proof process. The edit and proof steps are of course human subject matter expert linguists. The same internal risk management and quality control processes that you would use for a first pass human translator should then be applied. Some that have spoken out against the use of machine translation, both in the Life Sciences domain and other domains, often ignore the fact that the quality control processes are no laxer when using machine translation than when the first pass translation is performed by a human.

There are of course risks in using machine translation. What if the machine translation engine produces the wrong term? Is this not the same risk that is present with a human translator? In reality, if the data is managed well, actual metrics from Language Studio™ client projects has shown that there are fewer terminology and critical errors in output from a highly customized engine than made by a human translator. In many cases, raw Language Studio™ machine translation is delivering output that requires fewer edits than data produced by first pass human translators and many translation memories. Some customers have achieved productivity levels similar to that of an 85% fuzzy match on translation memories.

As with a human only approach, risks should be identified by analysis of the product and the risks involved in the incorrect use of the product. Whether the first pass translation was performed by a

HIGHLIGHTS

• Because of the quality level required for "mission-critical" translations, a rigorous quality control process is essential whether using human or machine translation. It is important to understand that the machine translation component is only used at the same stage as a first pass human translator.

• In many cases, raw Language Studio™ machine translation is delivering output that requires fewer edits than data produced by first pass human translators and many translation memories.

machine or a human, these risks are no different. There are many stages of human control after this very first stage. If the quality controls in place to catch such issues fail, then the same controls most likely would fail for a human first pass translator as well.

Of course, if you are not performing a full customization or working with older machine translation technologies or machine translation engines that are built with the Dirty Data SMT approach, then the risk is significantly increased. The risks are notably greater still if you are using a generic machine translation engine such as Google Translate or Microsoft Translator. Even a customized version built with Microsoft Translator Hub will typically be built on top of Microsoft's own data, which the user has no control over, so cannot be considered Clean Data SMT and will likely add a negative influence on the translation.

HIGHLIGHTS

• Using machine translation in the Life Sciences domain is no riskier than using a first pass translator for the same tasks if the customization of machine translation is performed correctly and the proper quality control processes are put into place.

• The very fact that the Life Sciences domain has very high-quality requirements makes it an ideal candidate for a fully customized machine translation engine. The Clean Data SMT model thrives on high quality data, consistent terminology and consistent writing style and delivers excellent output as a result.

• Machine translation is a productivity enhancer for human language experts. It is not a replacement and still requires the same validation and quality control that would be applied to a human translator or a translation memory.

CONCLUSIONS

- Using machine translation in the Life Sciences domain is no riskier than using a first pass translator for the same tasks if the customization of machine translation is performed correctly and the proper quality control processes are put into place. The only stage that changes when high quality and fully customized machine translation engines are deployed in the Life Sciences domain is that of how the initial translation is performed. Quality control is almost identical to that of human only approach. What is different is that the client's subject matter experts in localizing for the medical industry are involved in every step of customizing the machine translation engine and its ongoing quality improvement lifecycle.

- The very fact that the Life Sciences domain has very high-quality requirements makes it an ideal candidate for a fully customized machine translation engine. The Clean Data SMT model thrives on high quality data, consistent terminology and consistent writing style and delivers excellent output as a result. As such, companies involved in localizing Life Sciences content are better positioned than almost any other domain.

- Machine translation is a productivity enhancer for human language experts. It is not a replacement and still requires the same validation and quality control that would be applied to a human translator or a translation memory. Particular attention needs to be paid in areas where more than just translation is required and content must be adapted to the strict regulations for each target country.

Omniscien™ TECHNOLOGIES

HIGHLIGHTS

• A fully customized machine translation engine learns from the highest quality historical translations that have been processed through all of these most stringent quality control processes. As a result, the quality of the machine translation output is significantly higher than other domains. Nevertheless, it is recommended that quality control is maintained at the highest of standards.

• LSPs interested in trying machine translation in this domain should validate that the approach is based on the Clean Data SMT model and that they have the ability to control all of the data used in the custom machine translation engine.

- Both humans and machines will never be perfect. First pass human translators will make mistakes, even when they are subject matter experts. A fully customized machine translation engine learns from the highest quality historical translations that have been processed through all of these most stringent quality control processes. As a result, the quality of the machine translation output is significantly higher than other domains. Put simply, if you train a custom engine on high quality content that has been managed, refined and verified, you will get a better quality translation result. Likewise, if you train a custom engine on lower quality data, a lower quality translation result should be expected.

- Of course, we recommend that quality control is maintained at the highest of standards and that LSPs step carefully as they learn how to deploy machine translation, whether in the Life Sciences or any other domain. When using Language Studio™, we make this process much easier by providing the human cognition and guidance of Language Studio™ Linguists that engage directly with your subject matter experts to ensure the optimal quality possible. This level of empowerment by the end client means that the LSPs expert linguists are focusing on the important quality issues and not on the complex technical issues of customizing a high-quality machine translation engine.

When we talk to LSPs working in this domain, their primary issue is not about replacing human translators or about lowering costs, it is about having the ability to deliver to an ever-increasing demand.

A sizable portion of the content in this domain is mission-critical. As such, an appropriate level of caution is healthy and to be expected. LSPs interested in trying machine translation in this domain should validate that the approach is based on the Clean Data SMT model and that they **have the ability to control all of the data used in the custom machine translation engine.** Without this control and level of empowerment, the quality of the machine translation output will be considerably lower and the risk of errors will be higher. For this reason, we do not recommend any machine translation system in this domain irrespective of vendor that does not use the Clean Data SMT model.

As the domain has mission-critical content, we encourage LSPs to run an initial project that leverages machine translation as part of the workflow in parallel with the human only equivalent so that they can see for themselves the benefits and challenges that machine translation brings to the process. At the same time productivity metrics should be recorded for all steps of the process from the initial translation to the final deliverable to the end client so as to understand where more or less time is expended to deliver the same quality translation.

Omniscien™ TECHNOLOGIES

ABOUT OMNISCIEN TECHNOLOGIES

Omniscien Technologies is a leading global supplier of high performance and secure high-quality Language Processing, Machine Translation (MT) and Machine Learning technologies and services for content intensive applications. Our wide range of solutions serves clientele from various industries including the Localization Industry, Online Research Services, Publishing, eCommerce, Media, Online Travel, Technology, Enterprise and Government.

Omniscien Technologies has gained a reputation for cutting edge solutions with its Language Studio™ platform. Depending upon the customer's unique requirements, Language Studio™ can be deployed in a variety of ways to integrate with their in-house data processing and translation management systems, and it offers unparalleled levels of customization and control as well as feature rich pre- and post-processing, enabling customers with even the most complex data to achieve both high quality and high-volume output to satisfy every use case. Omniscien Technologies has by far the most comprehensive and feature rich system in the market today.

Covering 550 language pairs and with a number of industry specific solutions, Omniscien Technologies remains the partner of choice for customers with complex, high-volume bespoke data processing and machine translation needs.

For further information on Omniscien Technologies or Language Studio™, please visit **www.omniscien.com** or contact **sales@omniscien.com**

Omniscien™ TECHNOLOGIES

www.omniscien.com
sales@omniscien.com I +66 (2) 381 5869