Case Study: Omniscien Technologies and IOLAR

IOLAR was interested to build a custom machine translation engine to translate technical engineering content from German to Slovenian. This language pair has a relatively complex source language combined with a very difficult target language that, like other Slavic languages, has a large number of inflected forms.

While translator productivity was important, the primary objectives were to ensure a high level of writing-style consistency and terminological accuracy. As there was no specific and directly related translation memory available to train the system, several hundred thousand segments were gathered from several sources, in a much broader domain than technical engineering. This data was combined to form a single corpus that was used to train the engine.

Earlier Attempts with Moses

Based on the widespread publicity around Moses and the increasing number of public Moses Case Studies, IOLAR decided to try and use Moses to accomplish their machine translation objectives. Part of the decision to deploy Moses in-house was based around concerns over data privacy. Sharing data with a DIY Moses provider was a concern as many are also an LSP or tightly related to an LSP and may compete for the same business.

IOLAR invested six months of time in an attempt to build a Do-It-Yourself (DIY) Moses system of useable quality for a rather difficult language pair – German->Slovenian. A computational

linguistics expert was engaged and he spent three months building IOLAR's own custom engines using DIY Moses technology. At the end of the six-month period, IOLAR's Moses system was still producing unpredictable and unusable results. There were many problems with word order, terminology consistency, unknown words and incorrect inflected forms. Attempts made to understand and address the problems were unsuccessful.

IOLAR compared the output from their Moses engine with Google Translate output and found that Google still produced much better translation quality than their system. However, neither IOLAR's Moses engine nor Google Translate provided quality and productivity gains that would create any advantage to the business and many segments needed to be completely retranslated when post-editing was attempted. "Since our initial internal efforts did not progress with the desired speed we turned to Omniscien Technologies (formerly Asia Online) to deal with the growing urgency being communicated by our clients," said Simon Bratina, IOLAR's Executive Technical Director.

Custom Training Plan

Omniscien Technologies addressed IOLAR's data security concerns with a contract that provides comprehensive protection of the data that ensures that IOLAR maintains all the appropriate rights to the data and that Omniscien Technologies can only use the data for the purposes of customizing IOLAR's engines.

IOLAR provided the same translation memories that were used in their custom Moses engine for analysis and inclusion into the Language Studio™ custom engine, and worked with Language Studio™ Linguists to create a Customized Training Plan that addressed their specific goals. The Custom Training Plan identified issues and gaps in the training data and created a roadmap to address them.

Language Studio™ Linguists are specialist linguists that have had comprehensive training in the creation of commercially viable, high quality custom engines. The linguists, who possess very different skills to an NLP or computational linguistics specialist, focus on fine tuning engine data and algorithms to minimize post-editing efforts. NLP experts on the other hand, focus on the general science of language and computing, rather than on the real-world application and adaptation of data for commercial translation purposes. Language Studio™ Linguists use human cognition to determine which tools and automated processes will be applied to refine and create data to achieve the optimal results for a client – something that an automated process is not capable of today. A unique Custom Training Plan is developed for each custom engine, with a broad suite of data analysis and data manipulation tools used in conjunction with language and domain specific approaches to ensure optimal data preparation when building a custom engine. This differs considerably to the Do-It-Yourself (DIY) model where data is simply uploaded and trained.

While there were many issues that the Custom Training Plan addressed, four key issues were identified that would greatly increase translation quality and steps were added into the plan to address these issues:

Issue:	IOLAR's translation memories were from multiple sources and included mixed terminology for the same terms. This would result in inconsistent terminology in the translation output.
Solution:	In addition to the standard data cleaning that is part of the Clean Data SMT model, Language Studio™ tools were used to normalize terminology so that when translating the terminology choices were limited to those preferred by IOLAR.

Issue:	The domains that the translation memories originated from, while related, were not a match to the desired target domain of technical engineering. This resulted in many technical terms being unknown by the engine and significantly lowering the quality of translations. In Statistical Machine Translation (SMT) an unknown term can have a very negative impact on translation fluency and overall translation quality.
Solution:	Language Studio™ Advanced Data Manufacturing tools were used to perform gap analysis which identified several thousand unknown technical terms. Language Studio™ Advanced Data Manufacturing resolved the unknown terminology which were validated by IOLAR's linguists specialized in the domain.
Issue:	The writing style of the translation memories was mixed and not at all relevant to the target domain of technical engineering. Even if an understandable translation could be produced, it would be in the wrong context and style, and therefore need a large amount of editing in order to deliver publication quality.
Solution:	Language Studio™ Advanced Data Manufacturing tools were used to manufacture appropriate grammatical structures and contextual data in the correct writing style. This was driven by a deep analysis of the client's translation memories, and automated manufacturing of data that would adapt the writing style to the client's requirements.

Issue:	As Slovenian is a heavily inflected language, one of the very common issues was that the correct term was being translated, but in the incorrect inflected form. In many cases, the correct inflected form was not in the translation memories provided by IOLAR.
Solution:	Language Studio™ Advanced Data Manufacturing tools were used to manufacture appropriate inflected forms in the correct context. This data would be used to ensure that the correct inflected form was available in the training data and thus reducing the number of incorrect inflections in the output.

Initial Result

In contrast to IOLAR's Moses engine, the custom engine created with Language Studio™ was built quickly and without the need for specialized computational linguistics or NLP skills from IOLAR. This freed up IOLAR's translators to be able to work on more important tasks such as terminology refinement and validation. The resulting Version 1.0 engine was considerably better than IOLAR's previous internal efforts with Moses and was also higher quality that Google. While there was still plenty of room for improvement, this initial engine was useable for starting the pilot project.

Language Studio™ uses "Blind Test Sets" to measure initial quality using BLEU and other automated quality assessment metrics. Productivity metrics are used to validate the automated metrics. The initial Language Studio™ custom engine was 32 BLEU points better than Google Translate and 34 BLEU points better than Microsoft Translator. While BLEU is a useful indicator of quality, human productivity when post editing is a much better metric to indicate success, quality and value.

Quality Improvement Plan

Many of the error causing issues in a custom engine are not visible until the engine has been trained and the output can be inspected. This is particularly true of more complex language combinations such as German to Slovenian. The first version of a Language Studio custom engine is called a Diagnostic Engine for this reason. Much like the Custom Training Plan, the Quality Improvement Plan is based on a deep understanding of the specific issues that have been determined when Language Studio™ Linguists reviewed the output and data. Using their extensive experience in customizing thousands of translation engines, Language Studio™ Linguists created a plan specific to IOLAR's custom engine that delivered the most rapid improvement with the least effort.

In addition to the Quality Improvement Plan, Language Studio™ Linguists guided IOLAR through the post-editing process and showed them how initial post edited data could be fed back into the engine and used to quickly improve translation quality. Some training was also provided to IOLAR's team on how best to leverage runtime customization features in Language Studio™ such as Runtime Glossaries and Post Translation Adjustments which further improved quality and corrected some capitalization and formatting issues.

As the initial custom engine in its diagnostic release stage was good enough for the production usage test, the Quality Improvement Plan was able to incorporate valuable post editing feedback at this stage. While IOLAR was processing and post editing, Language Studio™ Linguists identified several improvement paths and manufactured additional data to improve grammar structures, word order and terminology consistency.

On receipt of the post edited data, analysis of the edits was performed by Language Studio™ Linguists again additional data was manufactured to reinforce the edits. These changes created an immediate 4 BLEU point increase that was validated by a noticeable increase in post editing productivity.

Conclusions

The IOLAR experience is an example of how a DIY approach might not work for production scale machine translation. During their "learning by doing" approach to DIY machine translation IOLAR spent a lot of time trying to understand why their initial efforts were producing such unpredictable results, and found that on some language combinations even the free online MT engines were easily outperforming their own Moses efforts.

The IOLAR example highlights *an inherent issue with DIY machine translation*, whether Moses based or from a commercial service – *it implies that the user knows how to do-it-themselves.* This case study demonstrates clearly that high quality machine translation requires considerably more effort, knowledge and skill than simply loading data into a system for training. Achieving a quality level that was usable for efficient post editing was clearly not the simple task that third-party DIY proponents had conveyed.

From a business perspective, it was clear that outsourcing to an expert was a better strategy than a DIY struggle, and I would say that our investment in Omniscien Technologies' (by that time Asia Online) Language Studio™ technology was one of the best technology investments that we have made.

…

Some of the very technical segments were the same quality as human translation.
– Simon Bratina,

Executive Technical Director, IOLAR

While some DIY Moses efforts are successful, few DIY Moses users know how to address or even identify the cause of problems when they do occur, even if they have some knowledge or

training in the core technological concepts. Moving beyond the initial problems in a DIY Moses custom engine is a significant challenge, even when expert NLP specialists or computational linguists were on staff. Skills in understanding data, not just algorithms and tools, are required to address the challenges in adapting, refining and creating data to address issues, either preemptively or as a remedy to issues.

Without a deep understanding of the cause of problematic machine translation output and corrective strategies to remedy them, the only improvement path available for most DIY Moses users is to upload post edited machine translations or additional translation memories. As there is little or no understanding of the impact that the new data will have, often the issues are not resolved and in many cases new issues and problems are introduced.

Language Studio™ Linguists provided IOLAR with the deep understanding of issues and provided efficient solutions to resolve critical issues affecting the quality of machine translation output. This ability to understand the data and error patterns has been gained through the creation of thousands of custom engines. Language Studio™ Linguists played a considerable part in taking this project from unsuccessful beginning on DIY Moses to being a considerable success in Language Studio™.

The overall conclusions and results drawn from IOLAR's collaboration with Omniscien Technologies:

- Working with an expert results in a much improved and significantly more efficient overall process.
- It is safer for an LSP to work with a non-LSP for technology that is as strategic as good MT can be.
- The long-term expertise and tools and capabilities like data manufacturing that Omniscien Technologies brought to bear on the process made it possible to reach high quality levels in just a few iterations.

- IOLAR noticed that there was a clear improvement in the machine translation output quality after the first iteration (incremental training) and they were surprised to see that “some segments were the same quality as human translation.”
- The data manufacturing and refinement tasks performed by Language Studio™ Linguists and further refined by IOLAR’s staff had greatly reduced the number of unknown words and incorrectly inflected forms, and delivered consistent terminology across translations.
- IOLAR achieved their core objectives of ensuring a consistent writing style and broad terminological accuracy that the clients had stated were of critical importance.
- IOLAR accomplished an improvement in the overall production efficiency.
- IOLAR realizes that while Moses may work for some simple cases where there is plentiful data in the target domain and language pair, deep expertise is required to produce successful systems outside of this atypical scenario. Even on these simple cases, it is now understood that with refinement of data along paths recommended by specialists, an even better result is possible.
- IOLAR is now making savings where it matters – in building the competences for an efficient post-editing when machine translation is used. While Moses is technically “*free*”, there are significant costs in staffing, hardware and other resources. There is also considerable risk in deploying a Moses system, even when hiring experts with NLP and computational linguistics experience.
- Even if IOLAR’s Moses system had delivered a quality that was better than Google, the savings and costs when compared to their investment in Language Studio™ would have been marginal. It has become clear to IOLAR that the Total Cost of Ownership (TCO) in a Language Studio™ system far exceed what was possible with DIY Moses solutions.

  A significant success factor of the collaboration between Omniscien Technologies and IOLAR is that IOLAR better understands the specifics of machine pre-translated text for difficult language combinations. IOLAR now grasps the post-editing requirements and necessary competences and will be able to approach customers that need such translations with confidence.

About Omniscien Technology

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 Localisation 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 548 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.

Contact:

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