Supporting Effective Intelligence Gathering with Statistical Machine Translation

TABLE OF CONTENTS

INTRODUCTION

Gathering intelligence from a corpus of high volumes of multi-lingual, multi-format content, possibly in near real-time, is a complex undertaking. However, recent advances in language processing technology and statistical machine translation (SMT) provide both the capacity (throughput) as well as the means to provide a rich set of information for analysis.

In different industry segments ranging from government to legal, there is an increasing need for effective machine supported analysis of multi-lingual and multi-formatted electronic data. In cases, such as eDiscovery, batches of data gathered need to be processed quickly and brought into a single language normalized format for further analysis while other uses include the monitoring of an ongoing (high-volume) stream of electronic data.

THE IMPORTANCE OF PRE-PROCESSING

More often than not, data under analysis is provided in a variety of different formats ranging from email to text documents, spreadsheets and presentations. Furthermore, the data is completely unstructured which makes effective analysis challenging at best.

Pre-processing data is key to “unlocking” the data and providing the basis for effective analysis. Pre-processing deals with the following key tasks:

1. **Normalization**
   Normalization is the process of bringing data into a uniform format, be it plain text or XML. By converting and optionally structuring the data, data can be processed more effectively and analyzed effectively by higher level tools.

2. **Translation**
   Having converted data into a uniform, machine readable format, the next important step is moving to a single language such that analysts can understand content and machine algorithms can be applied using a single language model.

3. **Tagging**
   As part of the statistical machine translation process complex language analysis is performed. This process provides valuable meta-data and tagging that is highly useful to the analysis stage. For example, so called “named entities” can be tagged accordingly which will enable the analytics process to in a later stage more effectively link related items.

4. **Enrichment**
   The pre-processing and translation stages can, based on the meta-data available, further enrich the data provided by adding additional tags and enrichment to the data. For example, locations identified can be tagged with GPS location information or people can be linked to online profiles and knowledge.

WHAT IS STATISTICAL MACHINE TRANSLATION (SMT) AND WHY USE A DEDICATED SYSTEM

Machine Translation technology has been known for years, however, early attempts were based around so called “rules based Machine Translation”. Rules based MT operated on the premise that it was possible to define clear rules for translation based on grammar and terminology that would enable a machine to translate effectively from one language to the other. While early systems existed that provide reasonable results for European languages in limited domains, applying this technology to a global pool of languages, all

HIGHLIGHTS

• Statistical Machine Translation operates on the basis of statistics, "learning" from different mono-lingual domain specific corpuses on how language is applied and then using bi-lingual aligned data to understand and ultimately perform the mapping (translation) from one language to the other.

• Different machine translation systems exist in the industry, however, control over the processing workflow and quality are of key importance to enable the technology to provide relevant business benefits.

possible language domains ranging from poetry to technology which apply very different writing styles quickly proved not to be feasible. Statistical Machine Translation on the other hand operates on the basis of statistics, "learning" from different mono-lingual domain specific corpuses on how language is applied and then using bi-lingual aligned data to understand and ultimately perform the mapping (translation) from one language to the other. Furthermore, SMT systems perform complex data pre- and post-processing to ensure the quality of the output is at the levels required.

Well known MT systems using SMT technology as a basis are Google Translate and Bing Translator to name two well-known examples and add rules where it enhances the result. However, while these systems do well for every day usage, they have three fundamental drawbacks in the context of intelligence gathering, namely:

1. **Data Privacy**
   The terms and conditions as well as the nature of systems such as Google Translate and Bing Translator mean that the respective organizations claim right to the data and have the ability to view the data which in most intelligence gathering or corporate environments is unacceptable.

2. **Quality**
   Generic MT engines such as the publicly available systems are trained with a wide range of variable quality content from the widest possible range of domains and do poorly in specialized content. In order to achieve high quality translations and reliably identify key entities, dedicated engines that have been trained on high quality data and extensive pre-processing and post-processing are required.

3. **Control & Scalability**
   Public systems provide no control over workflow, scalability or pre- and post-processing which are crucial to the quality and the meaningfulness of the output.

Machine translation (MT) has reached the stage of maturity and different systems exist in the industry. However, depending on the application scalability, control over the processing workflow and quality are of key importance in which case significant additional capabilities are required to enable the technology to provide relevant business benefits.

Omniscien™ TECHNOLOGIES

HIGHLIGHTS

• Machine Translation Systems enhanced with broader language processing capabilities provides extensive benefits for intelligence gathering by effectively identifying relevant and linked information from a corpus of data.

RELEVANCE TO INTELLIGENCE GATHERING

Intelligence gathering, be it in legal, government or other industries, is about quickly and effectively identifying relevant and linked information from a corpus of data. Machine Translation Systems enhanced with extensive language pre- and post-processing, such as Omniscien Technologies' Language Studio, can unlock this data. Language Studio™ provides extensive and scalable pre- and post-processing capabilities, translation and normalization. Tagging and enrichment can significantly improve the output quality and relevance while scaling as needed to even the highest volumes and reducing the lag between data being provided and relevant intelligence being gathered in an actionable form.

MT combined with broader language processing capabilities is highly relevant and beneficial to intelligence gathering in various industries and provides extensive competitive benefits both in terms of time to insight as well as overall capabilities

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