Beyond Text, Machine Translation and NLP for e-discovery

Jean Sennelart jean.senellart@systrangroup.com
SYSTRAN Global CTO/SYSTRAN SA CEO, Paris, 75002, France
Denis Gachot denis.gachot@systrangroup.com
SYSTRAN Software Inc President, San Diego, 92121, United States
Joshua Johanson joshua.johanson@systrangroup.com
SYSTRAN Software Inc, San Diego, 92121, United States

Abstract

As the amount and variety of digital data in different languages has increased, e-discovery processes need to evolve in order to streamline the processing of data and display crucial in-formation into an intuitive interface in a way that is scalable to any size user. It needs to be able to adapt to the user's needs and deal with any mix of media, such as image, voice, video, emails, blogs, social media posts and documents from any language in a manner that is con-sistent and intuitive. It needs to synthesize all of this information in a way that improves translation and exhibits the facts and evidences that the users need in the language of a digi-tal forensic examiner.

We will specifically describe how the tools implemented by SYSTRAN can be used to accomplish these more sophisticated tasks through several use cases. We will demonstrate how these tools can deal with multiple types of data, extract and normalize text, analyze lan-guage, create terminology lists, and customize the translation to better suit a variety of do-mains. We will illustrate how the tools accomplish this by self-tuning using unstructured and noisy corpora from the individual user and user-generated content written in approximate and sometime coded languages. This can be done across several languages, several types of data and multimodal documents, and can be scaled to suit the user's needs. We will show how these techniques can be used in combination to improve the overall translation quality and user experience.

Integration of SYSTRAN language libraries within an existing e-discovery platform will be presented to illustrate the presentation.

We will conclude by showing how these approaches can be generalized for big data analysis introducing challenges in real-time large scale data processing, but also processing of multi-topic and volatile information threads.

The full presentation can be found at http://static.systran.net/internaldocs/mtsummit-xv-systran.pdf