Datalambic

Semi-Automated Linguistic Data Acquisition

Realisation

Hieronymus AG

Paula Reichenberg

HES-SO Fribourg (iCoSys)

Prof. Dr. Jean Hennebert
Dr. Christophe Gisler
Donatien Burin-des-Rosiers

Keywords

- Contextual machine translation in law and finance
- Data collection based on pdf and web scraping
- Data pipelines with automated Machine Learning validation

Competences

Machine Learning
Deep Learning
Natural Language Processing (NLP)
Complex Data Systems

Valorisation

Commercialisation de LexMachina

Funding

Innosuisse
Application number:
48742.1 IP-ICT

Project duration

18 mois
11/2020 – 05/2022

Neural Machine Translation (NMT) engines have revolutionized the field of machine translation in just a few years. Translation engines such as Google Translate or DeepL have reached very fine performance levels on texts of a general nature. In specific contexts, translations are becoming increasingly technical and NMT engines need to undergo specialization through retraining based on contextualized data.

That is why **Hieronymus AG** has created LexMachina, the first NMT engine specializing in the translation of legal and financial texts. Its target users include law firms, banks, (re)insurance companies, consultants and the big4, in Switzerland and Germany. Such companies will benefit from higher quality custom-made NMT engines, with a shorter set-up time.

The purpose of the **Datalambic Project**, carried out in partnership with iCoSys, is to create a tool ecosystem for semi-automated collection, preparation and correction of high-quality data in order to (re)train neural translation engines in the desired specialization(s), including looped feedback from linguists, lawyers and users. The ecosystem will be modular, including: context-targeted web scraping, document classification, multilingual sentence alignment, term extraction, pseudonymization and customized machine translation post-processing.

1. Semi-automated data collection
- Seed-based web crawling and scraping
- *Language identification*
- *OCR, PDF / HTML conversion*
- *Text normalisation, tokenisation*

2. Data qualification
- Data collection optimizer:
  - topic verification, document pairing
  - crawl seed suggestions
- Semi automated phrase tag association

3. Data alignment
- *Bi-lingual alignment algorithms*
- Multi-lingual alignment, e.g. *LASER*
- Alignment quality evaluation
- Terminology alignment

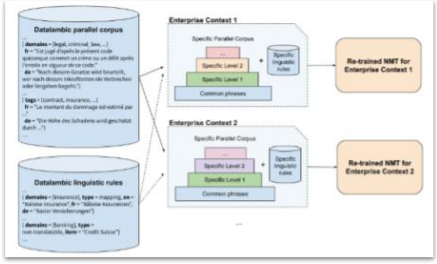
4. Corpus management tools
- Hierarchical corpus composition
- Bias detector
- Data collection orchestration and corpus composition traceability

5. Linguistic rule management
- NLP rule engine for text normalisation, NMT re-training or MT post-processing
- Automatic inference of rules, e.g. term extraction, term alignment

6. MT post processing
- *Post processing rule engine*
- MT quality check with alignment metrics
- Automatic selection of ensemble NMT
- *Anonymisation / pseudonymisation*

7. Active learning
- Human-in-the-loop active learning introduced in various components: crawl seed suggestions, topic verification, data alignment, linguistic rule discovery

8. Workflow composition and management
- Workflow engine to let linguists assemble sequences of components from WP1-6 for batch processing; saving and configuring such workflows



Retraining NMT engines for a given business context will require assembling corpora of parallel sentences using a pyramidal approach, for example: a set of "common segments" (e.g., sentences extracted from Swiss laws), supplemented by various sets of segments specific to increasingly sophisticated contexts (e.g., in the field of insurance, subsector of reinsurance, etc.)