INTERNATIONAL JOURNAL OF ENGLISH LANGUAGE, LITERATURE AND TRANSLATION STUDIES (IJELR)
A QUARTERLY, INDEXED, REFEREED AND PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
http://www.ijelr.in (Impact Factor: 5.9745) (ICI)

REVIEW ARTICLE

Vol. 8. Issue.2. 2021 (April-June)

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2628(Print):2349-9451(online)

DATA MINING AND RECENT ADVANCES IN THE STATE OF THE ART IN TRANSLATION MEMORY TECHNOLOGY

Dr. MANJUNATH AMBIG[1], V K S MAHALAKSHMI[2]

[1]Professor, Dakshin Bharat Hindi Prachar Sabha, Chennai

[2]Jr.Translator (OL), Electronics Corporation of India, Ltd (Department Of Atomic Energy), Hyderabad

V K S MAHALAKSHMI

Article information
Received:14/6/2021
Accepted: 28/6/2021
Published online:30/06/2021
doi: 10.33329/ijelr.8.2.250

ABSTRACT

In the current digital age in which we survive, information in the form of electronic data has been stored in large databases. Many companies have their own databases filled with stored data collected from either customers or saved information. One of the largest types of databases are those holding public data like, Technical Documents, Manuals, Annual reports, trademarks, Medical records, criminal records, licenses, legal documents, and many other. With the social media and web content/database is growing at exponential rates.

As technology is being upgrading day by day, and the value of electronic storage is becoming cheaper, more and more data and data is being stored electronically by corporations/companies and individuals around the globe, in many different languages. While it is helpful for a corporation to have a large massive data bank at its disposal, the ability to access and use that data is paramount to its value. Accomplishing this process is familiarly known as data mining, or text mining, is used.

Data mining also termed to as text mining, is the process of using software to search large volumes of data or information to extract out specific, related information for every one use. The better example of this technology would be using a search engine on the Internet, such as Google. If you were interested in information publicly openly available online on the topic of language translation, then you can go to the Google website and type in “language translation” into the search box and Google will return you a list of possible hits on that term based on its large bank of stored information which it has previously gathered. In this example, we are using data mining for “language translation.” Translation Software useful to understand foreign language content in a matter of seconds with the ability to customize the software to recognize specific terminology, phrases, and sentences in a detailed view Data mining translation allows you to search data stores, documents, and web pages for key terms and then translate the findings into Target language. Examples of text mining translation search unlimited items include, but are not limited to: trademarks, copyrights, patents, news articles, blogs, web pages, records databases, catalogs, and The Idealogy is to intergrate different accessible software tools for the

purpose of semiautomatic construction of Natural Language Ontologies (NLOs) from specific domains.

Keywords: Search box, Data base, Ontologies, Natural Language

Introduction

The current trending challenges facing the task of effective data mining today is the sheer volume of multilingual data that is stored. Suppose for a moment that you are a multinational corporation interested in text mining your corporation's data for a specific keyword or phrase for a report you are writing. Now suppose for a moment that your company stores its data in English, German and Japanese. While you can easily text mine for the English data, how do you search the German and Japanese data files without the ability to read or write in either of these two languages? The answer is simple – data mining translation using language translation software!

Data mining can be utilized to pull data from the internet or databases worldwide and then require the text be translated in order to search it further. For example there are companies that search public trademark databases in other countries when researching a product for development. Many times a search will produce thousands of documents that would be too time consuming for a human to look through, but with translation software you can program the software to segregate specific terminology that can be further researched and can be used.

Language Translation Solutions

In this paper, we explore the use of text mining techniques for translation memory maintenance. Language service providers often have large databases of translations, called translation memories, which have been in use for a long time leading to a slow population of the translation memory with other domains (i.e. adding financial content to a technical domain translation memory). To our best knowledge, no tools exist that would effectively separate the content of a translation memory according to different domains. Having the ability to extract individual domains from low-quality translation memories could mean a significant benefit to language service providers looking to utilize modern translation methods, such as machine translation and automated terminology management. In the first stage, we used OntoGen, a semi-automatic ontology building tool, to separate the segments in the translation memory according to domains. In the second stage, we wanted to test whether we could use OntoGen's topic keywords as shortcuts for building classification models– the reason for this being that manual annotation is costly and time consuming. If the topics extracted with OntoGen are accurate enough, then we could potentially skip the manual annotation phase of text classification, thereby significantly speeding up the process. We successfully managed to build ontology of the translation memory, but the boundaries between some topics were relatively vague. One reason for this is that we had to deal with sentences – as opposed to larger blocks of text – which are difficult to classify. Nevertheless, the results of the ontology creation were promising with manual evaluation showing that around 4 in 5 strings were assigned a correct label. The results of the second stage were less clear - the accuracy did significantly improve compared to the majority class classifier, but did not reach levels where it would be deemed useful in a professional language service provider environment.

Translation Memories (TM) are amongst the most utilizing tools by professional translators. The underlying idea of TMs is that a translator should benefit as much as possible from previous translations by being able to retrieve how a similar sentence was translated before. Despite the fact that the core idea of these systems relies on comparing segments (typically of sentence length) from the document to be translated with segments from previous translations, most of the existing TM systems hardly use any language processing for this. Instead of addressing this issue, most of the work on translation memories focused on improving the user experience by allowing processing of a variety of document formats, intuitive user interfaces, etc. explained in below figure 1.

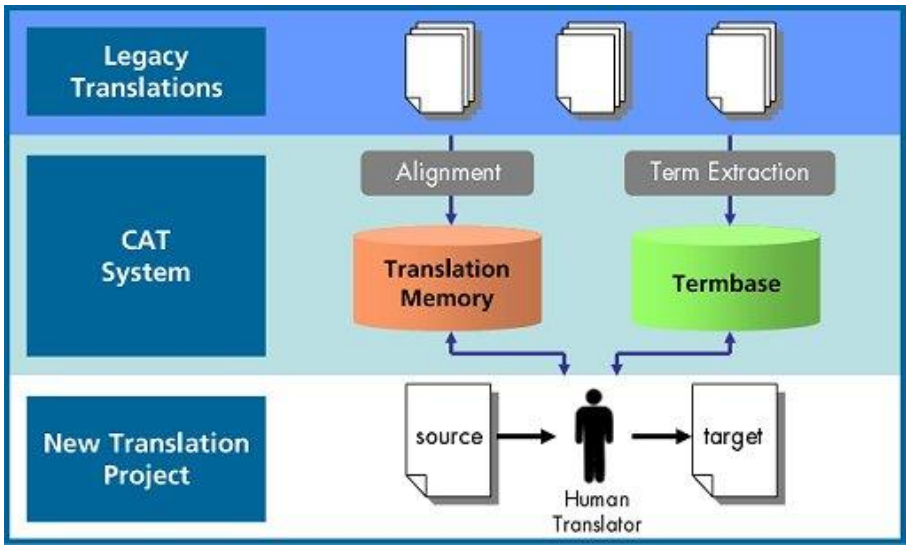

Figure 1.Translation Memory Management

Recent Advances in Google Translate

In the year 2016 Advances in machine learning (ML) have driven improvements to automated translation, including the GNMT neural translation model introduced in Translation studies that have enabled great improvements to the quality of translation for over 100 languages. Nevertheless, state-of-the-art systems lag significantly behind human performance in all but the most specific translation tasks. And while the research community has developed niche techniques that are successful for high-resource languages like Spanish and German and many other foreign languages for which there exist copious amounts of training data, performance on low-resource languages, like Yoruba or Malayalam, still leaves much to be desired. Many techniques have demonstrated significant gains for low-resource languages in controlled research settings (e.g., the WMT Evaluation Campaign), however these results on smaller, publicly available datasets may not easily transition to large, and web-crawled datasets.

In this paper, we share some recent progress we have made in translation quality for supported languages, especially for those that are low-resource, by synthesizing and expanding a variety of recent advances, and demonstrate how they can be applied at scale to noisy, web-mined data. These techniques span improvements to model architecture and training, improved treatment of noise in datasets, increased multilingual transfer learning through M4 modeling, and use of monolingual data.

Advances for Both High- and Low-Resource Languages Hybrid Model Architecture:

Four years ago we introduced the Recurrent Neural Networks RNN-based GNMT model, which yielded large quality improvements and enabled Translate to cover many more languages. Following our work decoupling different aspects of model performance, we have replaced the original GNMT system, instead training models with a transformer encoder and an RNN decoder, implemented in Lingvo (a Tensor Flow framework). Transformer models have been demonstrated to be generally more effective at machine translation than RNN models, but our work suggested that most of these quality gains were from the transformer encoder, and that the transformer decoder was not significantly better than the RNN decoder. Since the RNN decoder is much faster at inference time, we applied a variety of optimizations before coupling it with the transformer encoder. The resulting hybrid models are higher-quality, more stable in training, and exhibit lower latency.

Web Crawl: Neural Machine Translation (NMT) models are trained using examples of translated sentences and documents, which are typically collected from the public web. Compared to phrase-based machine translation, NMT has been found to be more sensitive to data quality. As such, we replaced the previous data collection

system with a new data miner that focuses more on precision than recall, which allows the collection of higher quality training data from the public web. Additionally, we switched the web crawler from a dictionary-based model to an embedding based model for 14 large language pairs, which increased the number of sentences collected by an average of 29 percent, without loss of precision.

Modeling Data Noise: Data with significant noise is not only redundant but also lowers the quality of models trained on it. In order to address data noise, we used our results on denoising NMT training to assign a score to every training example using preliminary models trained on noisy data and fine-tuned on clean data. We then treat training as a curriculum learning problem — the models start out training on all data, and then gradually train on smaller and cleaner subsets.

Pros of Low-Resource Languages in Particular: Back-Translation

In state-of-the-art machine translation systems, back-translation is especially helpful for low-resource languages, where parallel data is scarce. This technique augments parallel training data (where each sentence in one language is paired with its translation) with synthetic parallel data, where the sentences in one language are written by a human, but their translations have been generated by a neural translation model. **M4 Modeling:** A technique that has been especially helpful for low-resource languages has been M4, which uses a single, giant model to translate between all languages and English. This allows for transfer learning at a massive scale. As an example, a lower-resource language like Yiddish has the benefit of co-training with a wide array of other related Germanic languages (e.g., German, Dutch, Danish, etc.), as well as almost a hundred other languages that may not share a known linguistic connection, but may provide useful signal to the model.

In addition to general quality improvements, the new models show increased robustness to machine translation hallucination, a phenomenon in which models produce strange “translations” when given nonsense input. This is a common problem for models that have been trained on small amounts of data, and affects many low-resource languages. For example, when given the string of Telugu characters “ష ష ష ష ష ష ష ష ష ష ష ష ష ష ష ష”, the old model produced the nonsensical output “Shenzhen Shenzhen Shaw International Airport (SSH)”, seemingly trying to make sense of the sounds, whereas the new model correctly learns to transliterate this as “Sh sh sh sh sh sh sh sh sh sh sh sh sh sh sh sh”.

Conclusion:

The approach which explained to be general and applicable for any natural language through Data mining. The implied approach is very ambitious and clear because the problem NLO construction is inexact and tough. Understandably, certain adaptations and constraints are necessary depending on the features of the natural language in question. By incorporating back-translation into Google Translate, we can make use of the more abundant monolingual text data for low-resource languages on the web for training our models. This is especially helpful in increasing fluency of model output, which is an area in which low-resource translation models under performed.

Web & Books References:

[1]. https://www.scitepress.org/Papers/2019/73795/73795.pdf

[2]. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

[3]. https://towardsdatascience.com/introduction-to-machine-translation-9cb0e93e7cb

[4]. https://www.semanticscholar.org/paper/Example-based-machine-translation%3A-a-review-and-Hutchins/f09ed22d2689c89f29fbbeeff63a835c467e04ad

[5]. Somers, Harold L. 2001. Review article: Example-based machine translation. Machine Translation 14(14) 113–157

[6]. https://www.gala-global.org/knowledge-center/about-the-industry/language-technology