iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61193

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Using Machine Learning Methods for Evaluating Quality of Techincal Documents

Ankur
Delhi Global Institution of Technology, India

Abstract: ***In the context of an increasingly networked world, the availability of high quality transla-tions is critical for success in the context of the growing international competition. Large international companies as well as medium sized companies are required to provide well translated, high quality technical documentation for their customers not only to be suc- cessful in the market but also to meet legal regulations and to avoid lawsuits.***
Therefore, this thesis focuses on the evaluation of translation quality, specifically con- cerning technical documentation, and answers two central questions:
How can the translation quality of technical documents be evaluated, given the original document is available?
How can the translation quality of technical documents be evaluated, given the original document is not available?
These questions are answered using state-of-the-art machine learning algorithms and translation evaluation metrics in the context of a knowledge discovery process. The eval- uations are done on a sentence level and recombined on a document level by binarily clas- sifying sentences as automated translation and professional translation. The research is based on a database containing 22, 327 sentences and 32 translation evaluation attributes, which are used for optimizations of five different machine learning approaches. An op- timization process consisting of 795, 000 evaluations shows a prediction accuracy of up to 72.24% for the binary classification. Based on the developed sentence-based classifi- cation systems, documents are classified using recombination of the affiliated sentences and a framework for rating document quality is introduced. Therefore, the taken approach successfully creates a classification and evaluation system.

I. INTRODUCTION

A. Motivation

Being able to overcome language barriers is a topic of interest since the beginning of the17th century. Since then, devices and ideas have been developed to ease the understand-ing between people who speak different languages, such as mechanical dictionaries or universal languages. Due to highly internationalized companies and the overall global- ization, the ability to automatically translate texts from one language to another withouthuman assistance is a subject that has been addressed for roughly 60 years and has gainedmore interest throughout the last decade. To be successful in an international market,companies must provide well translated documentation for their products. As complexproducts often have more than one user group, e.g. administrators, users and developers, the number of different documents for a single product can be very high. Especially for export oriented companies, it is hard to find professional translators with an appropriate technical background to create properly translated technical documentation for reasonable costs. Therefore, machine translation solutions obtain increasing interest affected firms. It is of great interest, to provide automated, high quality translations of texts to ensure equal access to information regardless of their source language. In particular, accurate translations of technical documentation is a high priority for companies, since they are essential for a smooth workflow, customer satisfactions and describe the handling, func- tionality as well as security features of products. In this field, misunderstandings can be very severe. In addition, cooperation among companies can suffer from misunderstand- ings that are caused by bad translations.

Since natural language processing is a very complex issue, the output generated by ma- chine translation software still needs to be approved in order to ensure the required quality. Thereby, the simple use of software to translate business documents moves the problem from the creation of the document to the evaluation and correction, but does not solve it. Consequently, the evaluation of translated technical documentation is an important step for companies to reduce time and costs as well as to create an effective way of translating critical documents. Additionally, this ensures a certain level of quality. The difficulty in evaluating translation quality is due to the subjective nature and different aspects con- cerning the term quality, such as grammatical correctness, style improvements or semantic correctness.

©IJRASET: All Rights are Reserved

Having access to computerized systems that perform correct translations of any sentence is still visionary, especially due to the problem of translating the meaning of a sentence to the system. Concerning this problem, it is essential to be able to rank the qualityof a given translation - otherwise it is not possible to ensure that a document has been well translated. A focus on technical documentation is especially interesting, due to theirhigh quantity in every product selling company which further increases the motivationof automatically translating these types of documents. Nowadays, companies deal with the given translation problem of technical documents by outsourcing this task to external translators. Since the person requesting these translations does not necessarily speak the translated language, it is important to ensure that the work has been done properly and professionally by a human as ordered and not by an automated translation system.

Based on this background, the aim of this thesis is to select and implement a machine learning process that produces an algorithm, which is able to detect whether documents have been translated by humans or computerized systems. This algorithm builds the basic structure for an approach to evaluate these documents.

II. RELATED WORK

As mentioned above, the idea of automatically rating machine translations is not new. One of the basic methods discussed in the literature is the “round-trip translation”, whichworks by translating a text fragment into a foreign language and back. Afterwards, the original text and the newly generated text are compared [1]. On this basis, the “bilingualevaluation understudy” (BLEU) algorithm was developed and presented by Kishore Pa- pineni et al. in 2002. BLEU defines document quality as a strong correlation between machine translations and the work of professional human translators. It is based on the idea that in addition to word-accuracy, the length of a translated text is important. Accord-ing to Papineni et al., human translations have the tendency to be of higher quality and shorter than computerized translations [2]. This idea was further developed in the UnitedStates forming the *NIST algorithm for machine translation evaluation* by the National Institute of Standards and Technology. This algorithm weights matching words accord- ing to their frequency in the respective reference translation [3]. A second evolution of the BLEU metric, is the *Metric for Evaluation of Translation with Explicit Ordering*, called Meteor, which was developed by Lavie et al. in 2005. The main difference is Meteor’s ability to detect synonyms of words which results in potentially less erroneous translations [4]. Furthermore, Kulesza and Shieber (2004) propose the use of Support Vector Machines for classifying machine translations on a sentence level [5]. Extending on sentence level evaluation, there has been additional research on substituting the use of human-produced reference translations, which often come with high resource require-ments for the evaluation of machine translation systems. Popovic´ et al. propose the use oflexicon probabilities [6], while Gamon et al. suggest the use of pseudo references replac-ing the commonly used human reference translations with multiple automated translationsystems as references and combining calculated perplexity scores with a machine learn- ing classifier to evaluate sentence quality [7]. Finally, Albrecht and Hwa successfully useregression learning in combination with pseudo references [8, 9]. As shown above, a lot of research has been done on the topic of machine translation evaluation.

However, the focus on a specific domain of documents in order to gain implicit additional knowledge by using machine learning techniques is not sufficiently addressed and neitheris the comparison of different machine learning approaches in order to classify whether documents have been translated professionally or automatically. This work sets out to answer these questions.

III. PURPOSE AND RESEARCH QUESTION

In this thesis, a machine learning technique will be used in a knowledge discovery pro- cess to classify documents by their translation type (*professional translation*, *automatedtranslation*). Further, an approach on how to evaluate the quality of translated techni- cal documents will be proposed. Concerning this issue, we address two main research questions:

- How can the translation quality of technical documents be evaluated, given theoriginal document is available?
- How can the translation quality of technical documents be evaluated, given theoriginal document is not available?

IV. APPROACH AND METHODOLOGY

This work focuses on using machine learning methods and algorithms in order to evalu- ate translations of technical documentation. There are two different problems that will besolved within this thesis. First, translations of technical documents will be classified and evaluated with the machine learning algorithm having access to the original document. In the second attempt, an algorithm will be optimized on the same task without having knowledge of the original.

©IJRASET: All Rights are Reserved

The planned procedure for our master thesis is the following:
Based on research on existing methods and metrics, an iterative knowledge discovery process will be started to answer the given research questions. This process includes the determination of quality criteria for translated documents, the implementation of needed metrics and algorithms as well as the optimization of the machine learning approachesto solve the given task optimally. It is important to note that this process is of iterative nature, since the criteria and attributes as well as their impact on translation quality and classification possibilities will be determined by evaluating the algorithms' results using adatabase of technical documents and their translations. The used data set will range from automated translations of technical documents using computerized translation systemsto manual and professional translations. Furthermore, during this iterative process, the methods and algorithms used will be continually changed and optimized to achieve the best possible results. Finally, the process and results will be critically reviewed, evaluatedand compared to one another. The limits of automated translations with the current state of the art will be pointed out and a prospect for possible further developments and studieson this topic will be given.

V. SCOPE AND LIMITATION

Due to the fixed time frame, some limitations have to be set on this research to ensure thework can be finished in time.

1) The document classification and evaluation will focus on syntactic aspects of tech- nical documentation, while the semantic parts will be left out.
2) This work focuses specifically on technical documents. This focus has the potential to implicitly generate knowledge during the machine learning process, due to a smaller sized vocabulary compared to having no limitations on text domains. Other document domains will not be included in this thesis.
3) Since the examined technical documents did not provide multiple professional trans- lations, there will be no human references used for an evaluation of technical doc- uments. Instead, pseudo references, as described in section 3.2 will be used to circumvent the lack of human translations.
4) This work focuses on evaluations based on the results of machine learning ap- proaches. Other techniques, such as the creation of a new metric comparable to the BLEU or Meteor metric will not be taken into account.
5) This work will focus on translations between German and English.

VI. TARGET GROUP

First, this work is especially interesting for researches in the area of machine translation and machine translation evaluation by using combinations of different machine translation metrics and machine learning approaches. As mentioned in 1.1 the private interest group for this research are international companies, especially export oriented ones. This due to the fact that finding technically versed translators on a limited budget is clearly problematic. Second, this work is of interest for all kinds of customers, since it can lead to long-term improvement of available information for certain products. This can be of additional interest for customers and companies that are active in lesser populated languages.

Identification of the Data Mining Goal
The predefined goal of this thesis was to evaluate the quality of technical documents andtheir translations using machine learning methods with a focus on detecting the differ- ence between automated translations and professional translations done by humans. As mentioned in 1.3, the thesis addresses two research questions:

1) How can the translation quality of technical documents be evaluated, given the original document is available?
2) How can the translation quality of technical documents be evaluated, given the original document is not available?

To answer these questions, they were broken down into these practical working steps:

- Provide a machine learning algorithm with optimal prediction quality for identify-ing professional and automated translations of technical documents with and with-out access to the original document.
- Introduce a proper framework for ranking the quality of technical documents and fragments of technical documents regardless of their translation type.

This splitting was chosen, because the first steps including the data creation and pre- processing are rather similar among the two research questions. Therefore, it is more efficient to build a setup for providing a machine learning algorithm that solves the given data mining task with and without knowledge of the original document at once. The maindifference between the two variants lies in the choice of attributes, with substantially morevariables being allowed for the setup with knowledge.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The creation of a new framework for ranking the quality of technical documents will then be built on the outcome of our first working step. Therefore, the knowledge discovery process focuses on the first work-ing step and the second one is solved by using the achieved results.

An important limitation, as mentioned in section 1.5, is that this work focuses on the evaluation of syntax and does not take the semantic parts of a text into account, since thiswould go beyond the purpose of this thesis.

VII. CONCLUSIONS

This work answered the following research questions:

How can the translation quality of technical documents be evaluated, given the original document is available?

How can the translation quality of technical documents be evaluated, given the original document is not available?

This was done by using a knowledge discovery process consisting of the phases, gathering data, preprocessing data, choosing an appropriate data mining approach to find patterns among the data and interpreting them. Finally, the results were used for further research. The document database was broken down into a sentence level, producing nine data sets each containing 22, 327 data entries for each of two translation types (automated trans- lation and professional translation). 32 metrics and attributes were chosen and imple- mented, of which 18 needed a reference translation for the calculation process and 14 did not. To create a reference translation, one or two computerized translation systems were used respectively to translate the original document and generate a reference for the given candidate texts. The described data set was preprocessed, removing 5% outliers, and at- tributes correlating to one another with more than 90% as well as normalizing the data to generate comparable attribute values. The preprocessed data was used in multiple iter- ations for five machine learning algorithms, Decision Trees, Artificial Neural Networks, k-Nearest Neighbor, Naive Bayes and Support Vector Machines. The algorithms were op- timized on their parameters and tested on a holdout set that was split up from the database before training the models. The maximum results were achieved by the k-Nearest Neigh- bor classifier, scoring 72.24% while having access to the original document and 62.93% without access to it. To make a statement on document level classifications, the optimized algorithms were used on the sentences of each original document and the results were re- combined to classify the respective documentation. Having access to the original text, re- sulted in no classification errors for the 14 documents, while having no access to it showed a misclassification rate of 14.29%. To further validate the document-based results, a set of 19, 190 manufactured documents was created by randomly combining sentences to fictive documents of sizes varying from five to 3000 sentences. The observed misclassifications ranged from 34.40% for the smallest documents to no misclassifications for documents containing 250 or more sentences for algorithms trained without knowledge of the orig- inal document and from 26.90% for the smallest documents to no misclassifications for documents containing 100 or more sentences for algorithms trained with knowledge of the original document.

Furthermore, an evaluation framework was constructed to rank sentences and documents based on their quality regardless of their translation type. The proposed model consists of four classes using two optimized machine learning models for classifying the sen- tences and an additional reference independent grammar and spell check tool to generate a weighted mistake count for each sentence. In order to evaluate document quality, the quality classes of the respective sentences are averaged with additional weight for higher mistake counts.

Additionally, a translation database has been developed, containing 22000 professionally translated sentences. Their translation to German and further context information can be used for future research in the area of machine translation evaluation.

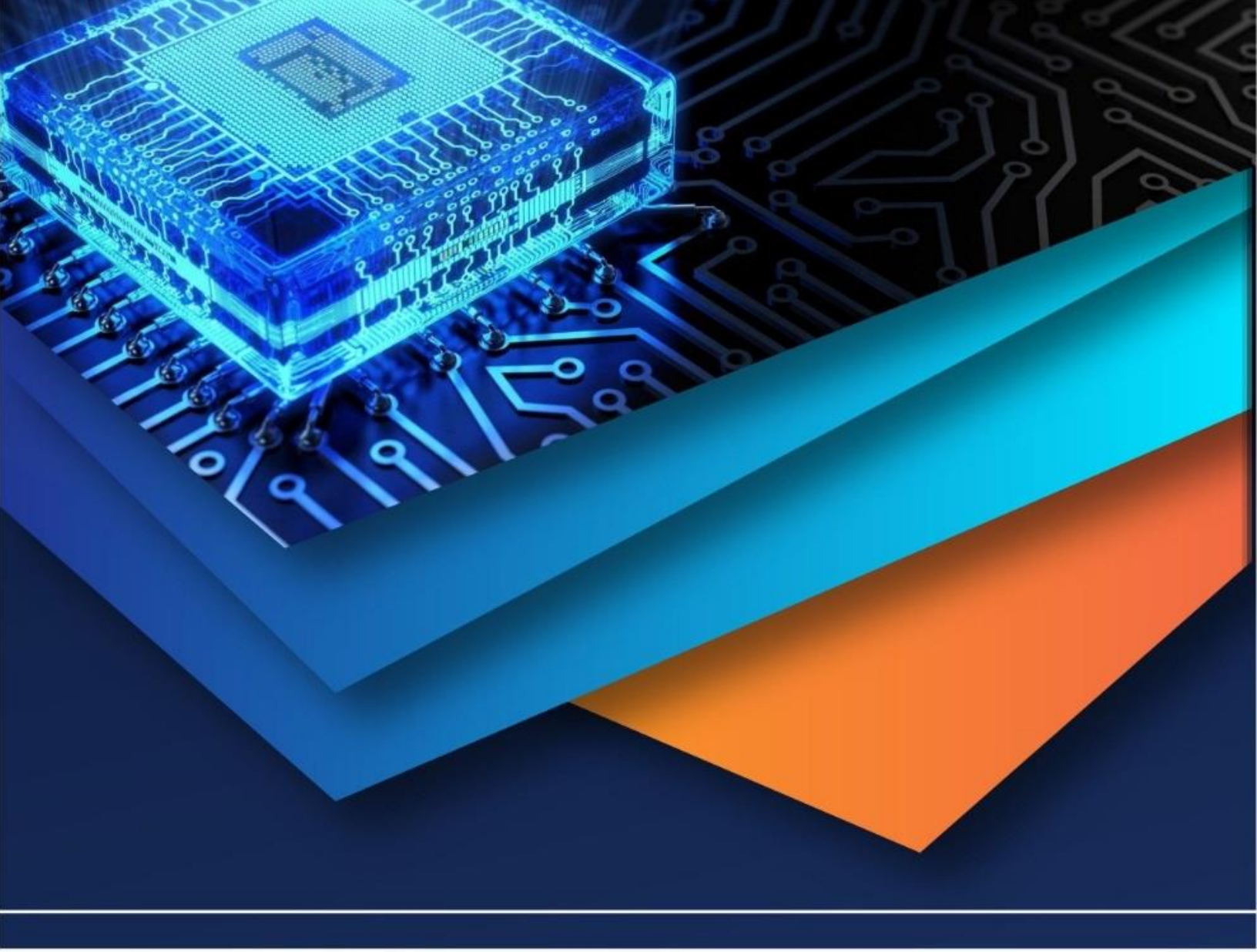
BIBLIOGRAPHY

[1] H. Somers, “Round-trip translation: what is it good for,” in In proceedings of the Australasian Language Technology Workshop, pp. 127–133, 2005.
[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Computational Linguistics (ACL), Proceed- ings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311–318, 2002.
[3] G. Doddington, “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics,” in Proceedings of the Second International Conference on Human Language Technology Research, HLT ’02, (San Francisco, CA, USA), pp. 138–145, Morgan Kaufmann Publishers Inc., 2002.
[4] A. Lavie and A. Agarwal, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” pp. 65–72, 2005.
[5] A. Kulesza and S. M. Shieber, “A learning approach to improving sentence-level mt evaluation,” in In Proceeding of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, 2004.
[6] M. Popović, D. Vilar, E. Avramidis, and A. Burchardt, “Evaluation without refer- ences: Ibm1 scores as evaluation metrics,” 2011.
[7] M. Gamon, A. Aue, and M. Smets, “Sentence-level mt evaluation without referencetranslations: Beyond language modeling,” in In European Association for MachineTranslation (EAMT), 2005.
[8] J. S. Albrecht and R. Hwa, “2007b. regression for sentence-level mt evaluation with pseudo references,” in In Proceedings of the 45th Annual Meeting of the Associationfor Computational Linguistics (ACL), pp. 296–303, 2007.

©IJRASET: All Rights are Reserved

[9] J. S. Albrecht and R. Hwa, "The role of pseudo references in mt evaluation," in In Proceedings of ACL, 2008.
[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," Ai Magazine, vol. 17, pp. 37–54, 1996.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)