# 3

# Predicting Errors in Google Translations of Online Health Information

MENG JI

## 3.1 Introduction

The use of machine translation in cross-lingual health communication and clinical settings is growing (Ragni and Viera, 2021; Manchanda and Grunin, 2020; Dew et al., 2018). Patients, medical professionals, and even common people with different language and cultural backgrounds have found these low-cost online translation systems very convenient. The technology can be especially useful for those with special needs, such as those with speech and hearing impairments. Overall, the use of online machine translations is on the rise across the world, but research shows that there are risks and uncertainties associated with these emerging technologies (Santy et al., 2019; Almagro et al., 2019; Mathur et al., 2013b; Kumar and Bansal, 2017). There is thus a pressing need to learn about the types and levels of mistakes and errors that machine translation systems make when deployed in health and medical domains. Policies and regulations are needed to reduce the risks and safety issues associated with the use of automated translation systems and mobile apps by clinicians and patients; and systematic empirical analyses of human and machine translation discrepancies of health and medical resources can inform their development.

Many online machine translation systems, such as Google Translate (GT), are constantly improving the quality of automated translation outputs by adapting technologies such as neural machine translation (NMT). Compared to traditional rule-based or statistical machine translation, NMT offers greater coherence, naturalness, and logical accuracy (Popel et al., 2020; Đerić, 2020; Jia et al., 2019), and is therefore more likely to gain trust from users who lack sufficient knowledge of the language pair being translated and consequently cannot judge the relevance and safety of translation outputs related to medical or health information. Research shows that several issues can in fact lead to serious errors in machine-translated health and medical resources.



https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

To be specific, the following features of English source texts were linked to clinically significant or life-threatening mistakes in machine translation outputs: (1) low readability of English long sentences (Flesch-Kincaid scores greater than Grade 8); (2) the use of atypical words, medical terminology, or abbreviations not explained in the source texts; (3) spelling and grammar anomalies; and (4) colloquial English (Khoong et al., 2019).

Nevertheless, despite its importance for reducing inequalities in healthcare services for vulnerable populations, improvement of translation technologies in medical and healthcare settings remains an understudied field of research.

In medicine and healthcare research, machine learning is becoming increasingly important. The detection and prediction of diseases, or of populations at risk of developing diseases, is an important application of machine learning, as early targeted interventions can improve the cost-effectiveness and efficiency of existing medical treatments significantly. The use of complex machine learning models can reduce investment in advanced medical experiments and clinical equipment but can also improve diagnostic precision by exploiting characteristics of the study subjects that are relatively easy to obtain. In general, classifiers that use machine learning tend to outperform the standard parameters and measurements of medical research in predicting health risks and diseases. It is true that the use of machine learning in health research has sometimes been criticized as overfitting learning algorithms due to small samples. However, some machine learning models, including sparse Bayesian classifiers such as relevance vector machines (RVMs), have proven to be highly effective in controlling algorithmic overfitting and thus improving the generality and applicability of findings (Madhukar et al., 2019; Langarizadeh and Moghbeli, 2016; Tipping 2001; Zhang and Ling, 2018; Silva and Ribeiro, 2006; Tipping and Faul, 2002).

This study examined whether it would be possible to improve diagnostic performance using Bayesian machine learning to combine easy-to-obtain English source health material features (both structural and semantic). It is anticipated that the results may lead to the automated combination and analysis of natural language features of English medical and health resources to improve detection of fundamental conceptual errors in translations into various languages. Success in detecting source text features associated with higher probabilities of conceptual errors in machine translation will support the use of machine learning techniques for the purpose of this study: the assessment and prediction of risk profiles of specific machine translations (Daems et al., 2017; Voita et al., 2019; Ashengo et al., 2021; Banerjee and Lavie, 2005).

For machine translations predicted to have high probabilities (>50 percent) of containing conceptual errors, a human evaluation and expert scrutiny would

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

be required to reduce potential risks and clinically significant errors for both users and communities. Translation error detection and prediction based on machine learning would improve awareness – among medical and health professionals and throughout the public at large – about how to safely use online translation software. In addition, this study examined the social implications of setting probability thresholds for Bayesian machine learning classifiers of machine translation error detection. Probability thresholds associated with higher classifier sensitivities and lower specificities imply higher predicted error rates in machine translation outputs; and these will result in increased investments in human review and greater burden on the healthcare systems of multicultural societies.

## 3.2 Methods

### 3.2.1 Research Hypothesis

As with human translation errors, conceptual errors in machine translation outputs can be predicted based on the likelihood of occurrence; and machine learning models can facilitate the prediction. For the purpose of this study, Bayesian machine learning classifiers were developed to predict the probability of critical conceptual mistakes (clinically misleading instructions) in outputs of state-of-the-art machine translation systems (Google). To develop the classifiers, the structural and semantic features of the original English source texts were used to estimate their risk profiles when submitted to machine translation tools online. The probabilistic output of sparse Bayesian machine learning classifiers is more intuitive for clinical use than machine learning output converted to nonlinear scales by postprocessing, and for this reason is more informative and preferable for the purpose of this study.

### 3.2.2 Screening Criteria for Text

MSD Manuals offer comprehensive medical resources developed by global health experts. Most of the original English sources have been translated into twelve world languages by professional translators and reviewed by domain experts since 2014. In China, these manuals are an important source of health education for the public and medical students (Liao et al., 2017). On the website of MSD Manuals' Consumer Edition, translated health resources are categorized by various common topics to facilitate search and retrieval of health information. Taking advantage of this resource for this study, 200 original

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

English texts were collected and, after removing texts not long enough for structural analysis, kept 185 articles of comparable lengths.

### 3.2.3 Topics of Infectious Diseases

With the aim of developing machine learning algorithms that are generalizable or topic-independent in predicting critical conceptual errors in Chinese machine translation, a cross-section of health resources on infectious diseases were selected. The collected texts related to the following diseases, among others: dengue, Ebola, Marburg virus, Hantavirus, hemorrhagic fevers, Lassa fever, lymphocytic choriomeningitis, Zika, bacteremia, botulism, clostridioides difficile infection, gas gangrene, tetanus, gram-negative bacteria such as brucellosis, campylobacter infections, cat-scratch disease, cholera, Escherichia Coli infections, haemophilus influenzae infections, klebsiella, Enterobacter, Serratia infections, legionella infections, pertussis, plague, Yersinia infections, pseudomonas infections, salmonella infections, shigellosis, tularaemia, typhoid fever, gram-positive bacteria such as anthrax, diphtheria, enterococcal infections, erysipelothricosis, listeriosis, nocardiosis, pneumococcal infections, staphylococcus aureus infections, streptococcal infections, toxic shock syndrome, and clostridioides difficile infection. Professional translators matched the original English texts with their Chinese translations, verified them in consultation with domain experts and published them on the Chinese edition of the MSD Manuals website.

### 3.2.4 Labeling of Machine Translations

Machine translations were generated using GT, using the original English source texts (May 2021). Chinese translations were labeled as human and machine translations respectively before being thoroughly analyzed by two native Chinese speakers trained as university researchers. They were asked to assess the severity of any discrepancy between paired Chinese translations (human versus machine). Language variability was allowed without causing clinically significant misunderstanding of original English source texts. A third trained observer adjudicated any discrepancies between the assessors. Machine translations exhibited two types of errors: terminological inconsistencies and conceptual errors. Conceptual errors were the focus of this study, in view of their higher severity and of the potential harm if machine-translated medical materials remained undetected by users lacking adequate medical training or the ability to appraise the materials.

### 3.2.5 Conceptual Mistakes in Machine Translations

In machine translation, conceptual mistakes are errors that can cause life-threatening actions or misinterpretations of original English materials. In this study on machine translation of public-oriented medical materials, these can include erroneous interpretation of medical advice, or clinical instructions on the detection, prevention, and treatment of infectious diseases and viruses. As an example, in an English text on preventing Ebola and Marburg virus infections, the original instruction was, "Do not handle items that may have come in contact with an infected person's blood or body fluids." Upon back-translation into English, the human translation closely matched the original meaning of the phrase, "Do not touch any objects that may have been contaminated with the blood or body fluids of the infected." However, the machine translation was "Do not dispose of objects which may have been touched by the blood or body fluids of the infected people." This discrepancy between human and machine translations was marked as a conceptual error since it was suspected that naive users of the machine translation output, lacking enough medical knowledge of the disease, might be unaware of the high risk of infection if they misunderstood the straightforward intent of the original medical instruction: not to clean or reuse Ebola patients' personal items.

In the same text, another critical, life-threatening conceptual mistake was found in the translation of the original English text "Avoiding contact with bats and primates (such as apes and monkeys) and not eating raw or inadequately cooked meat prepared from these animals." The human translation again matched the original meaning well: "Avoid touching bats and primates (like apes and monkeys) and not to eat the raw or not properly cooked meats of these animals." Machine translation by GT contained critical conceptual mistakes, as it read, "Avoid touching bats and primates (like apes and monkeys) and do not eat the raw and cooked meats of these animals." In another text on the prevention of Zika virus infection, the original text was, "Currently, men who may have been exposed to the Zika virus are not tested to determine whether they are infected and thus at risk of transmitting the virus through sexual intercourse. Instead measures to prevent transmission are recommended whenever people who may have been exposed to the Zika virus have sexual intercourse."

The human translation was close to the original meaning:

> Currently, men who may have been exposed to the Zika virus are not tested to confirm whether they are infected, as a result, the risk of getting infected through sextual intercourse exists. It is recommended that when having sex with men who may have been exposed to the virus, protective measures are taken to prevent infection.

When back-translated into English, Google's translation into Chinese meant, "Currently, men who may have been exposed to Zika virus are not tested to confirm whether they are infected, therefore there are risks of getting infected via sextual intercourse. By contrast, when having sex with men who probably have already been infected with the virus, protective measures are recommended to stop infection." The discrepancy in the Chinese translation of "whenever people who may have been exposed to the Zika virus" was marked as a critical conceptual mistake by machine translation, as the risk of virus infection via sexual transmission was clearly misinterpreted and downplayed. An ordinary Chinese user of machine translation might well be misled into believing that, as long as the individual has not been clinically diagnosed with Zika virus, it is safe to have sexual relations with that individual.

### 3.2.6 Prevalence of Conceptual Mistakes in Machine Translations

An extensive comparison of human and machine translations of the same English source text revealed similar conceptual mistakes in 89 texts (48 percent) of the total 185 texts collected for this study. In some cases, a machine-translated text contained as many as four or five conceptual errors. While the translation pair studied (English to Chinese) has been relatively well studied by machine translation researchers (including Google's), the high rate of persisting conceptual mistakes in machine translation of medical materials was alarming. Machine translation into and from less-researched languages is likely to generate higher rates of conceptual errors, especially for high-risk communities and populations speaking those languages.

### 3.2.7 Annotation of Features of English Source Texts

The English source texts were annotated with structural features using Readability Studio (Oleander Software). These features serve to quantify the morphological, syntactic, and logical complexity of original health materials in English. The following features are annotated: average paragraph length in sentences, number of difficult sentences (of more than twenty-two words), number of longest sentences, average sentence length in words, number of unique words, number of syllables, average number of characters per word, average number of syllables per word, number of proper nouns, number of monosyllabic words, number of unique monosyllabic words, number of complex (more than three syllable) words, number of unique multi-syllable (more

than three) words, number of long (more than six characters) words, number of unique long words, misspellings, overused words, wordy expressions, passive voice, and sentences beginning with conjunctions. In addition, to determine which words in English source texts are likely to cause conceptual errors when machine-translated into Chinese, the words of the original texts were annotated with their semantic categories, using the comprehensive automatic semantic tagging system developed by the University of Lancaster, USAS.

USAS contains twenty-one large semantic categories which are further divided into more than 100 sub-categories covering general and abstract words (A1–A15), the body and the individual (B1–B5), arts and crafts (C1), emotions (E1–E6), food and farming (F1–F4), government and the public (G1-G3), architecture, housing and home (H1–H5), money, commerce and industry (I1–I4), entertainment, sports and games (K1–K6), life and living things (L1–L3), movement, location, travel and transport (M1–M8), numbers and measurement (N1–N6), substances, materials, objects and equipment (O1–O4), education (P1), language and communication (Q1–Q4), social actions, states and processes (S1–S9), time (T1–T4), world and environment (W1–W5), psychological actions, states, process (X1–X9), science and technology (Y1–Y2), names, and grammars (Z1–Z99). These two sets of features are widely used in the development of machine learning models based on natural language processing techniques because they can be automatically annotated and interpreted relatively easily from the perspective of applied linguistics. In sum, in the final feature set, there were 20 structural features and 115 semantic features – composing a feature set sufficiently rich to enable exhaustive analysis and modeling of English source text features that may help to predict the occurrence of conceptual mistakes in the English-to-Chinese machine translation output.

### 3.2.8 Bayesian Machine Learning Classifier Relevance Vector Machine

The RVM is a variation of Support Vector Machines (SVM) (Cortes and Vapnik, 2005) which uses Bayesian inference and has the same functional form as SVMs (Tipping, 2001, 2004). As a Bayesian-based method, it offers probabilistic predictions and enables intuitive interpretations of uncertainty (Bishop and Tipping, 2003). The RVM model is also quite practical, in that it does not require large amounts of training data and generalizes well (Tipping, 2001; Bowd et al, 2008; Caesarendra et al., 2010). With these characteristics and advantages, it provides an ideal method for medical research and disease prediction. In these use cases, it is often necessary to determine the probability

of a disease based on observed symptoms, even though the relevant data is usually sparse and hard to collect (Bowd et al, 2008; Langarizadeh and Moghbeli, 2016). In this paper, an RVM model, enhanced by structural and semantic features, is applied to estimate the probability that machine translation of specific health education material s concerning infectious diseases will contain critical conceptual errors.

### 3.2.9 Training and Testing of Relevance Vector Machines with Three Different Full Feature Sets

In order to train and test machine learning classifiers, 70 percent of the data was used for training an RVM with three full feature sets, while 30 percent of the data was withheld for testing the three RVM models. The training data (129 texts in total) included 63 English source texts accompanied by machine translations containing conceptual errors, and 66 English source texts accompanied by machine translations without conceptual errors. There were 26 English source texts whose Chinese translations by GT contained conceptual mistakes, and 30 English source texts whose machine translations were correct. RVMs were trained using three feature sets to enable comparison of feature types: the full structural feature set (20); the full semantic feature set (115); and the combined feature set (135). To minimize bias in the classifier training process, five-fold cross-validation was applied to the training data (129). In particular, English source texts (the training data, 70 percent of the total data) linked or not with detected machine translation errors, were randomly divided into five approximately equal, exhaustive, and mutually exclusive subsets. Afterward, RVM classifiers were trained on four subsets combined and then tested on the fifth subset. The process was repeated five times, with each subset serving as the test data once. In this way, each tested English source text was never part of the training data and was only tested once. During cross-validation, a mean AUC (area under the curve receiver operating characteristic) and its standard deviation were calculated for the RVM trained on each full feature set. The remaining 30 percent of the testing data was used to evaluate the performance of the trained classifiers and to generate their sensitivity, specificity, accuracy, AUC, and F1.

### 3.2.10 Classifier Optimization

It was found that, in the current study, the large dimensionality (number of features: 135) and small sample size (185) of the data sets adversely affected the performance of the Bayesian RVM classifier in locating the separating

surface for classification. This classification uncertainty was reduced by using automated feature optimization to identify the best sets of structural and semantic features of the original English health texts, using backward feature elimination and 5-fold cross-validation to reduce bias in the optimized RVM classifier.

### 3.2.11 Backward Feature Elimination: RFE-SVM Method

Due to RVM's lack of "nuisance" parameters and its ability to automatically set regularization parameters to avoid overfitting, no hyper-parameter tuning was necessary to optimize the model (Tipping, 2001). To improve the performance of RVM, Recursive Feature Elimination (RFE) with SVM was applied as the base estimator (denoted as RFE-SVM) (Guyon et al., 2002) to reduce the feature dimension and automatically select the most important features that could improve RVM. For the RFE-SVM model, the parameter "min_features_to_select" (the minimum number of features to be selected) was set as "1" and set the "step" parameter (the number of features to be removed at each iteration) as "1." Z-score normalization was performed of the optimized features to improve the performance of the RVM classifier. As a result, the normalized data had zero mean and one unit deviation. The total set of health materials on infectious diseases was randomly split into training data (129) and test data (56) at a split rate of 0.7. The training data were used for feature optimization by 5-fold cross-validation and the performance of RVM with four different feature sets were evaluated on the remaining 30 percent test data. The cross-validation process of RVM classifier optimization was similar to the process used to train and test the full-dimension RVM classifier on the three feature sets (structural, semantic, and combined). First, the training data were divided into five subsets of approximately equal size. Four of the five subsets were used to determine the optimized feature set based on the maximum cross-validation score, using 5-fold cross- validation. The optimized feature set trained on the initial four subsets was then tested on the 5th subset to allow evaluation of the trained classifier with optimized features.

### 3.2.12 Separate and Joint Feature Optimization

The first step was to repeat the same process for the structural and semantic features separately, resulting in two separate optimized feature sets: the optimized structural feature set (OFT) and the optimized semantic feature (OSF) set. Features retained in the OFT set were as follows: average number

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

of sentences per paragraph, average number of characters per word, average number of syllables per word, passive voice, and sentences that begin with conjunctions. Features retained in the OSF set were these: expressions indicating probability (A7), possession (A9), food (F1), general substances and materials (O1), physical attributes (O4), speech acts (Q2), obligation and necessity (S6), power relationships (S7), time (general) (T1), time (beginning/ending) (T2), time (early/late) (T4), mental actions and processes (X2), sensory (X3), intention (X7), science and technology in general (Y1), and geographical names (Z2). Next, the two sets of separately optimized structural and semantic features were combined and labeled as "Combined Features via Separate Optimization" (CFSO) comprised of 21 features (5 optimized structural features and 16 optimized semantic features).

Lastly, the same feature optimization was repeated on the combined full feature set (135), using 5-fold cross- validation, yielding the "Combined Features through Joint Optimization" (CFJO: 48), a distinct optimized feature set with 11 structural and 37 semantic features. Structure and semantic features selected in separate optimization processes were quite different from those selected in the machine learning process, suggesting that the importance of individual features in machine learning depends largely on other optimized features. (Compare the situation in standard statistical analysis, where p values indicate whether variables are statistically significant.) The 11 structural features in CFJO were as follows: average number of sentences per paragraph, longest sentence, average number of characters, number of monosyllabic words, number of complex words of more than three syllables, number of unique multi-syllable words (more than three syllables), number of unique long words, misspellings, overused words, wordy items, and passive voice. The 37 semantic features in the CFJO set were: verbs/nouns indicating modify/change (A2), classification (A4), evaluation (A5), comparison (A6), probabilities (A7), possession (A9), degrees (A13), Anatomy and physiology (B1), health and diseases (B2), bravery and fear (E5), food (F1), furniture and household fittings (H5), life and living things (L1), numbers (N1), measurements (N3), quantities (N5), general substances/materials (O1), general objects (O2), linguistic actions, states, processes (Q1), speech acts (Q2), social actions, states, processes (S1); people (S2); obligation and necessity (S6); power relationship (S7); helping/hindering (S8); time: general (T1); time: beginning /ending (T2); time: old/new (T3); time: early/late (T4); sensory (X3); intention (X7); ability (X9); science/technology in general (Y1); geographical names (Z2); discourse connectors (Z4); grammatical expressions (Z5); and conditional expressions (Z7).

## 3.3 Results

The performance of RVM classifiers were compared using different optimized feature sets on the test data (Table 3.1, Figure 3.1): optimized structural features (OTF) (5), OSF (16), jointly optimized structural and semantic features (CFJO) (48), and separately optimized structural and semantic features (CFSO) (21). Table 3.1 shows that while the performance of optimized RVMs did not always improve over non-optimized RVMs on the *training* data (5-fold cross-validation), optimized RVMs were consistently much better than non-optimized RVMs on the *test* data. For example, AUCs of RVMs increased from 0.451 using original structural features to 0.587 using optimized structural features (OTF); AUCs of RVMs increased from 0.628 using original semantic features to 0.679 using OSF; AUCs of RVMs increased from 0.679 using original combined features to 0.689 using combined structural and semantic

Table 3.1 *Performance of RVMs with different feature sets on test dataset*

| | Training Data (5-fold CV) | Test data | | | | |
|---|---|---|---|---|---|---|
| Feature Sets | AUC Mean (SD) | AUC | Accuracy | Macro F1 | Sensitivity | Specificity |
| Original Combined Features (135) | 0.6166 (0.179) | 0.679 | 0.625 | 0.60 | 0.42 | 0.80 |
| Original Structural Feature (20) | 0.6319 (0.144) | 0.451 | 0.4821 | 0.48 | 0.54 | 0.53 |
| Original Semantic Features (115) | 0.6299 (0.166) | 0.628 | 0.6607 | 0.66 | 0.62 | 0.70 |
| Optimized structural features: OTF (5) | 0.6245 (0.078) | 0.587 | 0.5536 | 0.55 | 0.58 | 0.53 |
| Optimized semantic features: OSF (16) | 0.6837 (0.120) | 0.679 | 0.625 | 0.62 | 0.58 | 0.67 |
| Combined features through joint optimization: CFJO (48) | 0.6159 (0.105) | 0.689 | 0.6429 | 0.64 | 0.54 | 0.73 |
| **Combined features through separate optimization: CFSO (21)** | **0.6840 (0.111)** | **0.684** | **0.6786** | **0.68** | **0.73** | **0.63** |

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

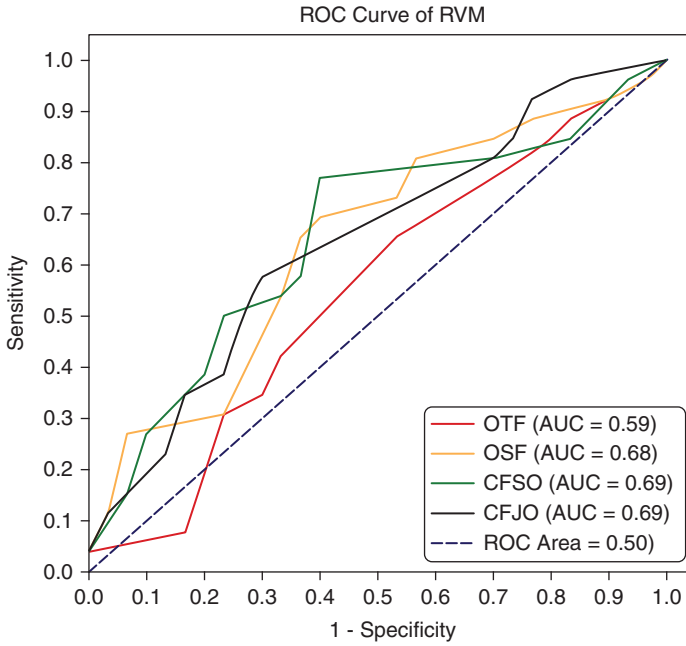


Figure 3.1 ROC curves of RVMs with different optimized feature sets

optimized features (CFJO). AUCs of RVMs did not improve using CFSO (0.684) over CFJO (0.689), but the total number of features was reduced by more than half from 48 (CFJO) to 21 (CFSO), and sensitivity of the RVM increased significantly from 0.54 using the CFJO features to 0.73 using the CFSO features. Specificity of RMV classifiers decreased from 0.73 using the CFJO features to 0.63 using the CFSO features. Since the goal of this study was to develop Bayesian machine learning classifiers that would detect and predict critical conceptual errors in machine translation outputs based on the observed features of the English source materials, higher sensitivity classifiers were deemed more useful for detecting mistakes in machine-translated Chinese health resources.

## 3.4 Comparison of Optimized RVMs with Binary Classifiers Using Readability Formula

This best-performing Bayesian RVM identified twenty-one features by separately optimizing structural and semantic features. The five optimized structural features were: average number of sentences per paragraph, average number of

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

characters per word, average number of syllables per word, passive voice, and sentences that begin with conjunctions. The sixteen optimized semantic feature were: expressions indicating probability (A7), possession (A9), food (F1), general substances and materials (O1), physical attributes (O4), speech acts (Q2), obligation and necessity (S6), power relationships (S7), time (general) (T1), time (beginning/ending) (T2), time (early/late) (T4), mental actions and processes (X2), sensory (X3), intention (X7), science and technology in general (Y1), and geographical names (Z2). The structural features included in the best-performing RMV resembled those incorporated in widely used readability formulas (Table 3.2). For example, the Flesch Reading Ease Score was based on average sentence length and average number of syllables per word; the Gunning Fog Index used average sentence length and percentage of hard words; and the SMOG Index used polysyllabic words (more than three syllables per word).

It was found the structural complexity of original English materials to have a significant impact on the quality of machine translation. In studying this relationship, the performance of the optimized RVM and binary classifiers was evaluated using some popular readability formulas (Flesch Reading Ease, Gunning Fog Index, SMOG Index) in terms of AUC, sensitivity, specificity, and whether the predictions of the optimized RMV and readability-formula-derived binary classifiers achieved statistically significant improvements over the reference line (AUC=0.5) (Table 3.3, Figure 3.2). The threshold of Flesch Reading Ease was 60, as texts with scores below 60 are considered fairly difficult to read, and texts with scores over 60 are easily understood by students ages 13 to 15. The threshold of SMOG Index and Gunning Fog Index was set at 12 to indicate a relatively easy reading level of medical texts in English, since scores above 12 tend to create reading difficulties and may increase the likelihood of conceptual errors in the machine translation output.

Table 3.2 *Readability formulas*

| Readability tools | Formulas |
|---|---|
| Flesch Reading Ease Score | Score=206.835-(1.015*ASL[a]) – (84.6*ASW[b]) |
| Gunning Fog Index | Score =0.4*(ASL[a]+PHW[c]) |
| SMOG Index | Score = 3 + Square Root of Polysyllable Count |

a ASL: average sentence length.
b ASW: average number of syllables per word.
c PHW: percentage of hard words.

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

Table 3.3 *Performance of the best-performing RVM with binary classifiers using readability formula*

| Test Result Variable and Thresholds | Area under the Curve | Std. Error[a] | Asymptotic Sig.[b] | Asymptotic 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound |
| **RVM (CFSO)** | **0.685** | **0.074** | **0.012 **** | **0.540** | **0.829** |
| SMOG (12) | 0.538 | 0.083 | 0.642 | 0.376 | 0.701 |
| Gunning Fog (12) | 0.533 | 0.080 | 0.677 | 0.376 | 0.690 |
| Flesch Reading Ease (60) | 0.492 | 0.082 | 0.925 | 0.333 | 0.652 |

a. Under the nonparametric assumption b. Null hypothesis: true area = 0.5

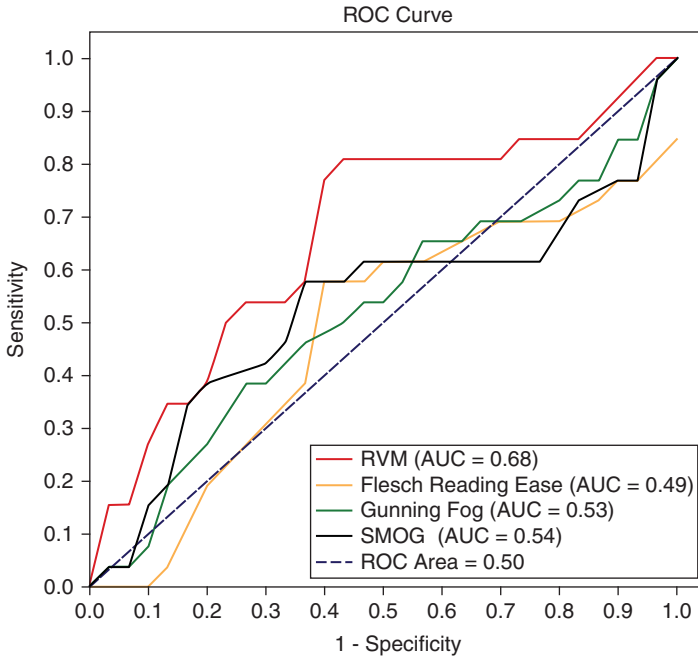


Figure 3.2 ROC curves of Flesch Reading Ease, Gunning Fog, and SMOG

Table 3.3 shows that AUC of the optimized RVM (using CFSO features) achieved statistically significant improvement of the reference (AUC= 0. 685, p=0.012, 95 percent confidence interval: 0.540, 0.829).

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

The three readability- formula derived binary classifiers did not improve over the reference (AUC=0.5): the AUCs of the SMOG Index and the Gunning Fox Index based classifiers were only slightly better than the threshold – respectively, 0.538 (p=0.642, 95 percent CI: 0.376, 0.701) and 0.533 (p=0.677, 95 percent CI: 0.376, 0.690); and the binary classifier using Flesch Reading Ease Scores was even less than effective than a random guess (AUC=0.492, p=0.925, 95 percent CI: 0.333, 0.652). Notably, according to this finding, the complexity of original English health materials, as measured by standard (currently available) readability parameters, cannot predict the presence of conceptual errors in machine-translated health and medical resources on infectious diseases. By contrast, however, a Bayesian machine learning classifier optimized based on the structural and semantic features of English input texts to the machine translation system did achieve statistically significant improvements in the prediction of conceptual mistakes in machine translation.

Table 3.4 shows the result of a pairwise resampled t test of the four classifiers: the optimized RVM and the three readability-formula based binary classifiers. It shows that although RVM achieved statistically significant improvement over the reference AUC (p=0.012), the improvement in AUC

Table 3.4 *Paired-sample area difference under the ROC curves*

| | Asymptotic | | | | Asymptotic 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| Test Result Pair(s) | z | Sig. (2-tail)[a] | AUC Difference | Std. Error Difference[b] | Lower Bound | Upper Bound |
| RVM vs. Flesch Reading Ease | 1.634 | 0.102 | 0.192 | 0.396 | −0.038 | 0.423 |
| RVM vs. Gunning Fog | 1.512 | 0.131 | 0.151 | 0.390 | −0.045 | 0.347 |
| RVM vs. SMOG | 1.466 | 0.143 | 0.146 | 0.393 | −0.049 | 0.342 |
| Flesch Reading Ease vs. Gunning Fog | −0.268 | 0.789 | −0.041 | 0.415 | −0.341 | 0.259 |
| Flesch Reading Ease vs. SMOG | −0.302 | 0.763 | −0.046 | 0.417 | −0.346 | 0.254 |
| Gunning Fog vs. SMOG | −0.131 | 0.895 | −0.005 | 0.389 | −0.082 | 0.071 |

a. Null hypothesis: true area difference = 0 b. Under the nonparametric assumption

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

was not statistically significant when compared with the three binary classifiers: the largest increase in AUC was between RVM and Flesch Reading Ease (0.192, p=0.102), followed by the AUC difference between RVM and Gunning Fog Index (0.151, p=0.131) and the AUC difference between RVM and SMOG Index (0.146, p=0.143). The AUC of the SMOG Index based classifier improved by 0.046 over the AUC of Flesch Reading Ease based classifier (p=0.763) and improved by

0.005 over the AUC of Gunning Fog Index based classifier (p=0.895).

### 3.4.1 Discussion

RVM produces probabilistic outputs through Bayesian inference, as opposed to SVMs. Bayesian probabilistic prediction enables relatively intuitive interpretation of classification results, and accordingly is relatively informative and helpful for clinical use and decision-making. According to this study, the best RVM classifier (AUC=0.685), based on two sets of separately optimized structural and semantic features, was able to usefully predict the probability that each specific original English text would belong to the group of texts associated with critical conceptual errors in machine-translated outputs. The RVM classified the original English text as a 'safe' text if its predicted probability was less than 50 percent, and as a 'dirty' text if its predicted probability was more than 50 percent. The RVM's probabilistic output gave an average mean probability of 0.388 (SD: 0.326, 95 percent CI: 0.266, 0.509) for 'safe' or error-proof English source texts and 0.606 (SD: 0.336, 95 percent CI: 0.472, 0.740) for 'risky' or error-prone English source texts.

Figure 3.3 is a histogram showing the percentage of English source texts in each 10 percent probability bin of the RVM probabilistic output for which conceptual errors in machine translations were detected (based on a comparison with human translations). 73 percent of the English source texts whose translations by Google contained critical conceptual errors were assigned a probability of "error-prone English text (EPET)" >= 50 percent (sensitivity: 0.73 percent); and 63 percent of English texts not linked with conceptual errors were assigned a probability of "non-error-prone English text (non- EPET)" > 50 percent (specificity: 63 percent).

The RVM results showed that most of the test English source texts associated with conceptual errors in machine translation belonged to the EPET group. (The distribution was negatively skewed, Figure 3.3.) For English source texts without conceptual mistakes in machine translation, the distribution of probabilities was less skewed. This result may be explained by the wide range of structural and semantic features of English source texts that are *not* related to

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

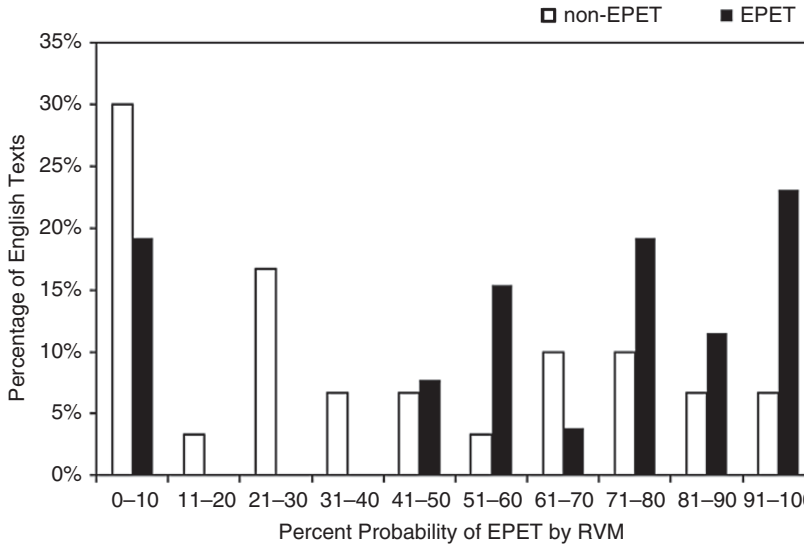


Figure 3.3 Percentage of non-mistake or mistake texts assigned by RVM classifier to 10 percent probability bins

conceptual errors in machine translation. 18 percent of the English source texts linked with machine translation errors were assigned to the 0–10 percent probability bin. This assignment indicated that there was still some uncertainty in the RVM probabilistic prediction, as some error-prone source texts were misclassified as "safe" source texts for machine translation systems.

Table 3.5 presents the various probability thresholds and associated sensitivity-specificity pairs of the best- performing RVM classifier, using a combination of structural and semantic features undergoing separate optimization. In real life, a meaningful probability threshold depends on the desired sensitivity-specificity pair. Classifiers of higher sensitivities are more suitable for screening purposes. Using the RVM, increasing numbers of English source texts were identified that would cause critical conceptual errors if translated using current machine translation tools, such as GT. However, increasing sensitivity can reduce specificity. And when specificities are lower, false-positive rates are higher (1-specificity), which means that more "safe" English source texts will be classified as error-prone or risky, even when the current translation technology can actually avoid life-threatening conceptual mistakes. And so, for health educational resource development and translation, lower screening classifier sensitivities imply heavier budgetary investments in human expert evaluation and assessment;

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

Table 3.5 *Under different probability thresholds, Sensitivity, Specificity and Positive Likelihood Ratios of the best-performing RVM with CFSO optimized features*

| SE-SP Pairs | Probability Cut-Offs | Sensitivity (SE) | Specificity (SP) | Positive Likelihood Ratio (LR+) |
|---|---|---|---|---|
| 1 | 0.075 | 0.846 | 0.300 | 1.209 |
| 2 | 0.415 | 0.808 | 0.600 | 2.019 |
| 3 | 0.494 | 0.769 | 0.633 | 2.098 |
| 4 | 0.496 | 0.769 | 0.633 | 2.098 |
| 5 | 0.586 | 0.577 | 0.633 | 1.573 |
| 6 | 0.625 | 0.577 | 0.667 | 1.731 |
| 7 | 0.703 | 0.5 | 0.767 | 2.143 |
| 8 | 0.757 | 0.385 | 0.800 | 1.923 |
| 9 | 0.799 | 0.346 | 0.833 | 2.077 |
| 10 | 0.876 | 0.269 | 0.900 | 2.692 |

and this issue in turn can result in further gaps in the provision of quality healthcare services and in support to populations and communities that rely on translated health resources and information for self-health management and disease prevention.

Another important indicator of the diagnostic utility of machine learning classifiers is the *positive likelihood ratio* (LR+), which is the ratio between sensitivity and false-positive (1-specificity) rates. Diagnostic utility increases with the positive likelihood ratio. In Table 3.5, sensitivity-specificity pairs (2, 3, and 4) showed high sensitivities (0.769–0.808) and moderate specificities (0.6–0.633), while positive likelihood ratios (2.019–2.098) showed small effects on post-test probabilities of English source texts causing critical conceptual errors in machine translations. The probability thresholds for these desirable sensitivity-specificity pairs (2, 3, and 4) were between 40 percent and 50 percent. As probability cut-offs increased over 50 percent, sensitivity decreased sharply, and specificity increased steadily. SE-SP pairs (5 and 6) produced the lowest positive likelihood ratios (1.573–1.731) and their probability thresholds were in the 50 percent-60 percent range. Finally, the pairs (7, 8, 9, and 10) were all impractical, as their sensitivities and specificities were very low, despite a positive likelihood ratio of 1.923–2.692. Since these models' sensitivities were low, they couldn't identify most English source texts that would likely result in critical conceptual errors if machine-translated using current systems. True, these high specificities did indicate that they were unlikely to over-predict the risk level of English source materials, thus requiring less expert evaluation and intervention, reducing healthcare budgets; however, in consequence, more

https://doi.org/10.1017/9781108938976.004 Published online by Cambridge University Press

clinically significant errors would be likely, because the screening classifiers would make professionals less aware of the high risks of using machine translation technologies clinically.

## 3.5 Conclusion

In cross-lingual health communication and clinical settings, machine translation is becoming increasingly common. It is true that these developing language technologies are associated with numerous risks and uncertainties, as research has shown. Still, in order to help reduce the risks of using such systems in clinical or patient settings, perhaps policies and regulations can be formulated based on evidence derived from systematic empirical analyses of discrepancies between human and machine translations of health and medical resources. With this goal in mind, the present study has sought to determine the probabilistic distribution of mistakes in neural machine translations of public-oriented online health resources on infectious diseases and viruses, using as predictors various linguistic and textual features that characterize English health-oriented educational materials. Two-hundred English-language source texts on infectious diseases and their human translations into Chinese were obtained from HON.Net-certified websites on health education. Native Chinese speakers compared human translations with machine translations (GT) to identify critical conceptual errors.

To overcome overfitting problems in machine learning for small, high-dimensional data sets while aiming to identify possible source text features associated with clinically significant translation errors, Bayesian classifiers (RVM) were trained on language-specific source texts classified as yielding, or not yielding, machine translation outputs containing critical conceptual grammatical errors. Among the best-performing models, the RVM trained on the CFSO (16 percent of the original combined features) performed best. RVM (CFSO) outperformed binary classifiers (BCs) using standard English readability tests. The accuracy, sensitivity, specificity of the three BCs were as follows: FRE (accuracy 0.457; sensitivity 0.903, specificity 0.011); GFI (accuracy 0.5735; sensitivity 0.685, specificity 0.462); and SMOG (accuracy 0.568; sensitivity 0.674, specificity 0.462).

In this study, Bayesian machine learning classifiers with combined optimized features did in fact identify certain features of English health materials features as associated with (and possibly causing) critical conceptual errors in state-of-the-art machine translation systems. It was found that machine-generated Chinese medical translation errors were most associated with certain

English structures (e.g., passive voice or sentences beginning with conjunctions) and semantic polysemy (different meanings of the same word when used in different contexts), since these features tend to lead to critical conceptual errors in NMT systems (English to Chinese) of health education information on infectious diseases. This finding challenges the hypothesis that complex medical terminology and low linguistic readability are the main causes of critical translation errors, since none of the predictor features appeared to be related to these factors.

Overall, this study underlines the need for clinical and health education settings to be cautious and informed when using the latest translational technology. It also points toward provision of helpful aids in exercising that caution. Classifiers can be trained using machine learning models like ours to identify texts containing features likely to yield clinically significant translation errors. Tools found to cause more such errors for the same texts could be avoided. At the same time, recommendations could be made for preemptively revising the original source texts to minimize likely errors. Additionally, machine learning might be applied to *automatically* revise source texts. Finally, the findings and procedures might be used to augment existing confidence scores for real-time translations, so that users could be warned that a current translation was suspect, and that paraphrase might be advisable.

## Reference

Almagro, M., Martínez, R., Montalvo, S., & Fresno, V. (2019). A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation. Journal of Biomedical Informatics, 94, 103207. https://doi.org/10.1016/j.jbi.2019.103207.

Ashengo, Y. A., R. T. Aga, and S. L. Abebe. 2021. "Context Based Machine Translation with Recurrent Neural Network for English–Amharic Translation," *Machine Translation*, 35 (1), pages 19–36. https://doi.org/10.1007/s10590-021-09262-4.

Banerjee, S., and Lavie, A. 2005. "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics, Ann Arbor, MI.

Bishop, C. M., & Tipping, M. E. 2003. "Bayesian regression and classification," *NATO Science Series sub Series III Computer And Systems Sciences*, 190, pages 267–288.

Bowd, C., J. Hao, I. M. Tavares, et al. 2008. "Bayesian Machine Learning Classifiers for Combining Structural and Functional Measurements to Classify Healthy and Glaucomatous Eyes," *Investigative Ophthalmology and Visual Science*, 49 (3), 945–953.

Caesarendra, W., Widodo, A., and Yang, B. S. 2010. "Application of Relevance Vector Machine and Logistic Regression for Machine Degradation Assessment," *Mechanical Systems and Signal Processing*, 24 (4), pages 1161–1171.

Cortes, C., & Vapnik, V. 1995. "Support vector machine," *Machine Learning*, 20 (3), pages 273–297.

Daems, J., Vandepitte, S., Hartsuiker, R. J., and Macken, L. 2017. "Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort," *Frontiers in Psychology*, 8. https://doi.org/10.3389/fpsyg.2017.01282.

Đerić, I. 2020. Google Translate Accuracy Evaluation, in Sinteza https://doi.org/10.15308/Sinteza-2020-80-85. http://portal.sinteza.singidunum.ac.rs/paper/745.

Dew, K. N., A. M. Turner, Y. K. Choi, A. Bosold, and K. Kirchhoff. 2018. "Development of Machine Translation Technology for Assisting Health Communication: A Systematic Review," *Journal of Biomedical Informatics*, 85, pages 56–67, https://doi.org/10.1016/j.jbi.2018.07.018.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. 2002. "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, 46 (1) pages 389–422.

Jia, Y., Carl, M., and Wang, X. 2019. "Post-editing Neural Machine Translation versus Phrase-Based Machine Translation for English–Chinese." *Machine Translation*, 33 (1), pages 9–29.

Khoong, E. C., Steinbrook, E., Brown, C., and Fernandez, A. 2019. "Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions." *JAMA Internal Medicine*, 179 (4), pages 580–582.

Kumar, A., and Bansal, N. 2017. Machine translation survey for Punjabi and Urdu languages. In 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall) (pages 1–11). IEEE.

Langarizadeh, M., and Moghbeli, F. 2016. "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review." *Acta Informatica Medica*, 24 (5), 364. https://doi.org/10.5455/aim.2016.24.364-369.

Liao XP, Chipenda-Dansokho S, Lewin A, Abdelouahab N, Wei SQ. 2017. "Advanced Neonatal Medicine in China: A National Baseline Database," *PLOS ONE*, 12(1): e0169970. https://doi.org/10.1371/journal.pone.0169970

Madhukar, N. S., Khade, P. K., Huang, L., et al. 2019. "A Bayesian Machine Learning Approach for Drug Target Identification Using Diverse Data Types." *Nature Communications*, 10 (1), pages 1–14.

Manchanda, S., and G. Grunin. 2020. Domain Informed Neural Machine Translation: Developing Translation Services for Healthcare Enterprise, In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisbon, Portugal, European Association for Machine Translation. www.aclweb.org/anthology/2020.eamt-1.27.

Mathur, P., Ruiz, N., and Federico, M. 2013b. "Recommending Machine Translation Output to Translators by Estimating Translation Effort: A Case Study." *Polibits*, 47, pages 47–53. https://doi.org/10.17562/pb-47-5.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., and Žabokrtský, Z. 2020. "Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals." *Nature Communications*, 11 (1), pages 1–15.

Ragni, V. and L. N. Vieira. 2021 "What has changed with neural machine translation? A critical review of human factors." *Perspectives*, 30, (1), pages 1–22.

Santy, S., S. Dandapat, M. Choudhury, and K. Bali. 2019. Interactive Neural Machine Translation Prediction, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Hong Kong, Association for Computational Linguistics.

Silva, C. and B. Ribeiro. 2006. "Scaling Text Classification with Relevance Vector Machines," *Systems Man and Cybernetics. SMC'06*, 5, pages 4186–4191.

Tipping, M. E. 2000. "The Relevance Vector Machine." In *Advances in neural information processing systems*, pages 652–658.

Tipping, M. E. 2001. "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, 1 (Jun), 211–244.

Tipping, A. and A. Faul. 2002. "Analysis of Sparse Bayesian Learning," *Advances in Neural Information Processing Systems*, 14, pages 383–389.

Voita, E., R. Sennrich, and I. Titov. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion, In arXiv:1905.05979v2

Zhang, Y., and C. Ling. 2018. "A Strategy to Apply Machine Learning to Small Datasets in Materials Science," *NPJ Computational Materials*, 4 (1). https://doi.org/10.1038/s41524-018-0081-z