Automatic Selection of Parallel Data for Machine Translation

Despoina Mouratidis, Katia Lida Kermanidis

Automatic Selection of Parallel Data for Machine Translation

Despoina Mouratidis [0000-0002-2844-5488] and Katia Lida Kermanidis [0000-0002-3270-5078]

Department of Informatics, Ionian University, 7 Pl. Tsirigoti, 49100 Corfu, Greece
{c12mour, kerman}@ionio.gr

Abstract. Nowadays machine translation is widely used, but the required data for training, tuning and testing a machine translation engine is often not sufficient or not useful. The automatic selection of data that are qualitatively appropriate for building translation models can help improve translation accuracy. In this paper, we used a large parallel corpus of educational video lecture subtitles as well as text posted by students and lecturers on the course fora. The text is quite challenging to translate due to the scientific domains involved and its informal genre. We applied a random forest classification schema on the output of three machine translation models (one based on statistical machine translation and two on neural machine translation) in order to automatically identify the best output. The unorthodox language phenomena observed as well as the rich-in-terminology scientific domains addressed in the educational video lectures, the language-independent nature of the approach, and the tackled three-class classification problem constitute innovative challenges of the work described herein.

Keywords: Machine Learning·Educational data·Data selection·Machine Translation·Random Forests.

1 Introduction

In recent years, many people, companies and organizations make use of machine translation (MT) solutions. MT software has been improving, and researchers are trying to generate the best translation of a source text. The use of MT is said to have become an indispensable tool, not only for scientific purposes, but also for the general public. Moreover, automatic translation contributes decisively to the learning process, since it can extend the learning target group by breaking the language barrier and enhancing access to the educational material. To this end, the European project TraMOOC (Translation for Massive Open Online Courses) [10] aims at improving the translation process, and overcoming the language barriers in online educational content.

After almost half a century that statistical approach prevailed in MT [9], a new method, the neural – based approach, appeared. This, in contrast to statistical machine translation (SMT) implemented by using parallel text corpora to calculate probabilities, generates much more accurate translations. More specifically, neural machine

translation (NMT) implements deep learning techniques to teach text translations by taking on existing statistical models as a basis. Also, NMT is able to use algorithms to train itself with linguistic rules [3].

Certainly, many challenges occur in the translation process. Additionally, it should be noted that there is difficulty in translation out of domain data. Therefore, there is a large amount of data to be translated. For large sentences, an extension of the classical neural encoder-decoder can be used, taking into account only the words which have information relevant to the target word and not the whole sentence [1]. Post-editing and data selection are two ways to reduce the data (i.e. choosing only quality parallel segments) without hurting translation quality. Many online MT platforms now prompt users to improve the proposed translation themselves [12]. This may be a solution to improve translation models, but it also creates a multitude of data that needs to be evaluated regarding usability. On the other hand, data selection methods can be used to recognize the useful and non-useful features in parallel segments. Research has shown that when models are trained with less, but more accurate, data, their performance improves [17].

In this paper, we consider data selection as a classification problem and we explore the idea of using three translation prototypes for our experiments, one based on SMT and the other two based on NMT. The contribution of this paper is multi-fold:

- the educational content domain comprises scientific fields that involve a high degree of terminology and unknown words. This phenomenon requires a set of robust learning features to represent the parallel text segments.
- the informal genre (spontaneous speech transcriptions and forum text) presents linguistic phenomena that are unorthodox and ungrammatical, like repetitions, interjections, fillers, truncated utterances etc., posing a challenge to the automatic identification of grammatical utterances.
- the proposed approach is language independent. All linguistic features are based on string similarity and no morphosyntactic information is incorporated in any form.
- a metalearner (Random Forest) is employed for data classification, for the first time for the task at hand to the authors' knowledge, in order to tackle the aforementioned challenges.

2 Experimental Setup

This section describes the corpora, tools and the classification process used.

2.1 Corpus

The parallel corpus we worked on was provided from the TraMOOC project. As already mentioned, the corpus includes lecture speech transcriptions and text posted by class participants on course fora. The source corpus consists of 2,687 segments (sentences) in English (Src). For each of these segments, three translation outputs into

Greek are available, generated by three prototypes (Trans1, Trans2, Trans3), whereas one reference translation (Ref) from a professional translator is also provided. Translation model 1 (Trans1) used the open-source phrase-based SMT toolkit Moses [8], the translation models 2 (Trans2) and 3 (Trans3) used the NMT Nematus toolkit [15]. Trans1 is a statistical based prototype trained on both in- and out-of-domain data. Trans2 is trained on the same data as Trans1 and uses labels to identify and remember the domain, while Trans3 is the result of training with more in-domain data providing via crowdsourcing, weight tying, layer normalization, and improved domain adaptation. Out of domain data included widely known corpora e.g. Europal, JRC-Acquis, OPUS, WMT News corpora etc. In domain data included TED, QED corpus, Coursera etc. [11].

Data pre-processing included the removal of symbols (for example #, $), and some alignment corrections, so that each segment is mapped to its Src, Ref, Trans1, Trans2 and Trans3 variations.

A challenge was the translation of entities like URLs, mathematical expressions and rare words. The first two entity types were copied to the translation output by some prototypes (Trans2 & Trans3, and Trans3 respectively), while the third type is tackled by the third prototype by word division in order to improve MT output [16].

2.2 Annotation

Two Greek linguists have annotated each segment with A, B or C depending on whether Trans1, 2 or 3 is more similar to Ref respectively. We observe low annotation percentage for class A (17%) compared with class B (37%) and C (46%). This confirms the superiority of the NMT vs. SMT models. At this point, it's important to notice that the two annotators gave different answers in 82 of the 2,687 segment cases. For the different answers, the annotators had a discussion and finally agreed on one common label.

We present five segments and their Trans1-2-3 and Ref translations (Table 1), the sixth segment is an example of disagreement of two linguistics.

Table 1. Segment Examples from source, Trans1, Trans2, Trans3 and Ref.

ID	Source	Trans1	Trans2	Trans3	Ref
1	The archplot refers to the classical design of a story and has been called by many names.	Η archplot αναφέρεται στην κλασική σχεδιασμό μιας ιστορίας και έχει κληθεί με πολλά ονόματα.	Το αρχαϊκό σχέδιο αναφέρεται στον κλασικό σχεδιασμό μιας ιστορίας και έχει κληθεί από πολλά ονόματα.	Η αρχική πλοκή αναφέρεται στον κλασσικό σχεδιασμό μιας ιστορίας και έχει ονομαστεί από πολλά ονόματα.	Η κύρια πλοκή αναφέρεται στην τυπική διαμόρφωση μιας ιστορίας και έχει πάρει πολλά ονόματα.
2	A bit of gaming his-	Ένα κομμάτι της ιστορίας	Ένα κομμάτι της ιστορίας	Ένα κομμάτι ιστορίας παι-	Λίγη ιστορία παιχνι-

	tory: Which now famous video game character made his/her first appearance in the 1981 "Donkey Kong" arcade game?	παιχνιδιών: Η οποία τώρα διάσημο βιντεοπαιχνίδι χαρακτήρας έκανε την πρώτη της εμφάνιση στο 1981 "Donkey-Kong" βιντεοπαιχνίδι;	του παιχνιδιού: Που τώρα ο διάσημος videoπαιχνίδι χαρακτήρας έκανε την πρώτη του εμφάνιση το 1981 στο βιντεοπαιχνίδι του Doney-Kong;	χνιδιών: Το οποίο τώρα ο διάσημος χαρακτήρας του βιντεοπαιχνιδιού έκανε την πρώτη του εμφάνιση στο παιχνίδι "DonkeyKong" παιχνίδι;	διών: Ποιος/α σημερινός γνωστός χαρακτήρας βιντεοπαιχνιδιού εμφανίστηκε για πρώτη φορά το 1981 στο «Donkey Kong»;
3	This is where studying Critical Thinking can help.	Εδώ είναι που σπουδάζουν σειρά μαθημάτων Κριτική Σκέψη μπορεί να βοηθήσει.	Εδώ είναι που η μελέτη της κρίσιμης σκέψης μπορεί να βοηθήσει.	Εδώ είναι που η μελέτη της Κριτικής Σκέψης μπορεί να βοηθήσει.	Σε αυτό το σημείο οι σπουδές στην Κριτική Σκέψη μπορούν να βοηθήσουν.
4	Upload the essay as a zip file including the Statement of Authorship.	Ανέβασε το δοκίμιο ως ταχυδρομικό φάκελο, συμπεριλαμβανομένης της δήλωσης Niemann.	Ανέβαζε την εργασία ως φερμουάρ, συμπεριλαμβανομένης της δήλωσης του αρχαίου πλοίου.	Ανεβάστε την έκθεση ως ένα αρχείο zip συμπεριλαμβανομένου της δήλωσης του διατάκτη.	Ανεβάστε την έκθεση ως συμπιεσμένο αρχείο συμπεριλαμβανομένης της Δήλωση Συγγραφικής Πατρότητας.
5	You need to get the audience to want to "lean into the screen".	Θα πρέπει να πάρετε το κοινό να θέλουν να "λιτή στην οθόνη".	Πρέπει να κάνεις το κοινό να θέλει να "λυγίσει στην οθόνη".	Πρέπει να κάνεις το κοινό να θέλει να "γείρει στην οθόνη".	Θα πρέπει να κάνετε το κοινό να θέλει να «μπει στην οθόνη».
6	For anybody interested in deeper exploration of the origins of storytelling please check-out Professor	Για όποιον ενδιαφέρεται για βαθύτερη εξερεύνηση του προέλευση της αφήγησης παρακαλώ ελέγξτε-	Για οποιονδήποτε ενδιαφέρεται για βαθύτερη έρευνα για την προέλευση της αφήγησης, παρακαλώ εξετάστε την	Για οποιονδήποτε ενδιαφέρεται για βαθύτερη εξερεύνηση της καταγωγής της αφήγησης, παρακαλώ ελέγξτε την	Όποιος ενδιαφέρεται για πιο διεξοδική έρευνα σχετικά με την προέλευση της αφήγησης

	Hobohm's full lecture on the topic that we added below.	out ο καθηγητής Hobohm είναι γεμάτο διάλεξη για το θέμα που προσθέσαμε παρακάτω.	πλήρη διάλεξη του καθηγητή Hobohm για το θέμα που προσθέσαμε από κάτω.	πλήρη διάλεξη του καθηγητή Χόμπομ για το θέμα που προσθέσαμε παρακάτω.	παρακαλώ ρίξτε μια ματιά σε όλη τη διάλεξη του καθηγητή Hobohm πάνω στο θέμα που προσθέσαμε από κάτω.

ID 1: i) *archplot*: No translation by Trans1 (not found). Trans2 and Trans3 correctly separate the two synthetics. Trans2 translates the first synthetic as a main word (*αρχαϊκό=archaic*), common in historical contexts, but not correct in this segment. Trans3 finds the meaning of the prefix: arch- (archi › αρχή, αρχική).

ii) *has been called*: the three trans didn't change the passive into the active form. Trans1 and 2 gave the most common meaning (*κληθεί*). Nevertheless, the more successful translation of Trans3 (*ονομαστεί*) makes a pleonasm with the object (*ονόματα*), Ref's choice being the correct (*πάρει*).

ID 2: i) *which*: None of the three Trans translated correctly this question word, not being the first word of the segment.

ii) Trans1 incorrectly connected *which* to history giving the same grammatical gender (*ιστορία...η οποία*). Trans2 chose the sometimes confusing, but very common, *που* (not the question word πού). Trans3 incorrectly connected *which* to *a bit* giving the same grammatical gender (*κομμάτι...το οποίο*).

iii) *now*: None of the three Trans translated it as an adjective.

iv) Trans1 and Trans2 didn't connect the word *game* to *character* as a genitive case (*παιχνίδι...χαρακτήρας*), as was done correctly by Trans3 (*χαρακτήρας βιντεοπαιχνιδιού*).

ID 3: i) *This is where*: no metaphorical sense by the three Trans.

ii) *studying*: the same translation by Trans2 and Trans3 (*μελέτη*), but not expressing the action, the process, as a verb would have done. Trans1 uses a verb (*σπουδάζουν*) and the sense of "process" is given also by adding an object but the syntax generates a pleonasm (*σπουδάζουν σειρά μαθημάτων*) and the syntax of the segment is incorrect.

ID 4: i) *Authorship*: Trans1 translated this word by the word *Niemann* that is non-existing in the source segment. It's important to note that we find the *Niemann Statement* in Harvard and other contexts and this is relevant with essays and authors. Very interesting, (but the result is completely false) is also the Trans2 translation process: from the basic meanings of the whole word (*authorship*= origin, source) Trans2 uses a synonym: *αρχαίου* (ancient), but at the same time it separately translates the second synthetic of the word (*-ship*) to give the common phrase: *αρχαίου πλοίου*.

ii) *essay*: Trans1 gives the main meaning of the word (*δοκίμιο*), but the Trans2 and Trans3 choices are also correct (*εργασία, έκθεση*), Trans3 choice being Ref's choice.

iii) *zip file*: Trans1 translates *zip* by the common adjective of *file*, but here it is irrelevant: *ταχυδρομικό*. The Trans2 translation (*φερμουάρ*) is completely irrelevant here, but very common in other contexts. Trans3 correctly doesn't translate *zip* in this context.

ID 5: i) *get*: Trans1 gives the most common translation (*πάρετε*), but it is not correct here. Trans2 and Trans3 correctly translate this multi-sense word.

ii) *You*: Only Trans1 correctly translates *You* as a plural pronoun.

iii) *lean*: None of the three Trans is correct *(λιτή, λυγίσει, γείρει*) compared to Ref's correct choice (*μπει*). The word here has a very special metaphorical meaning: "to enter". The sense of "motion" of the preposition *into*, in the Source text, is partly conveyed in Trans3 (*γείρει*).

ID 6: Annotator 1 labeled Trans2 as the better translation for the following reasons: 1) *έρευνα* is a better translation for *exploration* in this segment, as the main meaning of the word (*εξερεύνηση*) here is not precise, 2) *προέλευση* for the word *origins* is the best translation in this segment and is the same as the Ref tranlation, 2) *εξετάστε* is not the best translation for *check-out* in this segment but is better than *ελέγξτε*, because its meaning is not the primary one (i.e. check) but closer to other secondary meanings of *check*, like "note" or "hold". By *check-out*, in combination with the word *whole*, the writer here means: "read" or better "study and keep in mind", but it can't be translated so, as it is far from the meaning of *check*, 3) Trans2 kept the proper noun *Hobohm* in Latin letters, like Ref, and as it is considered to be good practice for dealing with proper nouns from one language to another.

Annotator 2 labeled Trans3, as the better translation for the following reasons: 1) *εξερεύνηση* is the exact meaning of "exploration", in combination with its prepositional phrase *of the origins*, implying "deeper research" (*εξερεύνηση* being more exploratory than a simple research (*έρευνα*)), 2) *της καταγωγής* is the primary and most common meaning for *of the origins*, as it refers to "the first appearance", to "the creation" of the subjective genitive: *storytelling*, 3) Trans3 changed the Latin into Greek letters for the proper noun: *Χόμπομ*, as the target language is Greek, and it is common practice to do so.

2.3 Features

We considered the task at hand as a classification problem with three output (class) values, so we represented each segment as a tuple (Src, Trans1-2-3, Ref). Each tuple was modeled as a feature-value vector, while the features are based on string similarity, they contain no form of morhosyntactic information, and are therefore language independent. The feature set was based on the work by Barron-Cedeno et al. [2] and Pighin and May [13]. Feature values were calculated using MATLAB.

Basic-Simple Features

These are simple string similarity features. Levenshtein distance is a string similarity metric, which calculates the minimum number of single-character changes re-

quired to change one word into the other. Also, another string similarity metric was used to determine if Trans 1-2-3 is contained in Ref (Containment c) [5].

- Length (in number of words) of Src-Trans1-Trans2-Trans3-Ref.
- Length in words of Trans1, Trans2, Trans3, Ref divided by Src, also for Trans1, Trans2, Trans3 divided by Ref.
- Length in characters divided by length in words for Src, Trans1, Trans2, Trans3 and Ref.
- Levenshtein distance of Trans1, Trans2, Trans3 divided by Length of words and characters of Trans1, Trans2, Trans3.
- Number of words that exist in Trans and do not exist in the Ref divided by the number of words in Trans (and vice versa).
- Containment c of Trans1-Ref, Trasn2-Ref, Trans3-Ref [5].
- Ratio of (third bullet)'s resulting features between (Trans1-Trans2-Trans3, Src), (Ref, Src), and (Trans1-Trans2-Trans3, Ref).
- Longest word for Src, Trans1, Trans2, Trans3 and Ref.
- Longest word in Trans1, Trans2, Trans3 divided by Src and Ref and longest word in Ref divided by longest word in Src.
- If Ref=Trans1 or Trans2 or Trans3, then True, otherwise False, if Src = Trans1 or Trans2 or Trans3, then True, otherwise False.

Noise-based Features

- If Src is a one word string then True, otherwise False.
- If Src is a string of more than five words then True, otherwise False.
- If Src is a string with length six to ten words then True, otherwise False.
- If Src is a string with length up to eleven words then True, otherwise False.
- If Src, Trans1, Trans2, Trans3, Ref has a word with length 10 to 14 characters then True otherwise False. We did the same with word length 15 to infinity.
- If Src or Trans1 or Trans2 or Trans3 has a word of three repeated characters then True, otherwise False.

Similarity-based Features

- The length factors (LF-defined in [14]), LF(Ref, Trans1), LF(Ref, Trans2) and LF(Ref, Trans3) are calculated.
- Using the LF (described above), if LF(Ref, Trans1)>LF(Ref, Tans2) then True, otherwise False. The same comparison is performed on LF(Ref, Trans2) and LF(Ref, Trans3), as well as on LF(Ref, Trans1) and LF(Ref, Trans3) (and vice versa).

2.4 Results

We have nominal and numeric features. We normalized the numeric features so that their values range between 0 and 1, by using the Feature scaling method. We decided

to use the Weka machine learning workbench [18] for training and testing our dataset. We used evaluation measures that are common in classification, and adopted from Information Retrieval. The first measure is Precision, that is True Positive / (True Positive + False Positive). The second measure is Sensitivity - (Recall), that is True Positive / (True Positive + False Negative).

Given the challenges governing the genre and the domain of the data, we decided to apply a meta-learner for increased robustness. We chose the Random Forest classifier, due to their using the Law of Large Numbers and the ability to avoid overfitting [4], and achieving high generalization accuracy. Random Forests implement an ensemble learning schema that generates multiple decision trees during training, and constructs a combination of the classification outputs of each tree model for prediction. We set the number of iterations (number of trees to be constructed) to 65. Each tree was constructed while considering 20 random features. We employed 10 fold cross validation as testing mode. The minority class (A) causes problems in the classification process: the classification algorithms give low accuracy as they tend to classify the new unseen segments in the majority class [7]. In order to improve the accuracy of the classifier for the minority class (precision 49%, recall 22%), we used the Smote filter [6], which is an over-sampling approach for creating new synthetic training data. Smote combines the feature values of minority class examples with the feature values of their nearest neighbor examples (n=5) in order to produce new examples of the minority class. The Smote process is applied only on the training data. Using Smote, the segments of class A doubled in number, and the total number of segments reached 3150. We observed that we had better results when we used RandomForest_Smote including all the features as seen in Table 2.

Table 2. Precision and Recall of our experiments.

Classifier : RandomForest				
Class	Precision	Recall	Number of features	Number of instances
A	49%	22%	82	2687
B	46%	36%	82	2687
C	50%	70%	82	2687
Classifier : RandomForest_Smote				
A	77%	63%	82	3150
B	44%	32%	82	3150
C	50%	68%	82	3150

It is noted that the results obtained are satisfactory, given that in our experiment we had three classification values, in contrast to related research that targeted a binary class output [2]. Moreover, the features we used are simple string comparison features, and they are language independent, including no morphosyntactic information in any form.

It is noted that the results obtained are satisfactory, given that in our experiment we had three classification values, in contrast to the [2] research. In addition, the features we used are simple string comparison features, and they are language independent.

We observed in the table that when we applied RandomForest before the Smote filtering, the classifier correctly classified 22% of A segments for Class A (Trans1), 36% of B segments for Class B (Trans2) and 70% of C segments for Class C (Trans3). After the Smote process, a major change is observed in Class A, where the percentage increased to 64%. For classes B and C we did not notice any particular changes. What is remarkable is that when the classifier does not sort correctly, it usually classifies the segments from one neural model to another (60% B -> C and 25% C -> B), and a much smaller percentage to the statistical model (7% C -> A and 8% B -> A) as well. In total, we can see in the figure below (Fig. 1) the percentages of incorrectly classified instances.

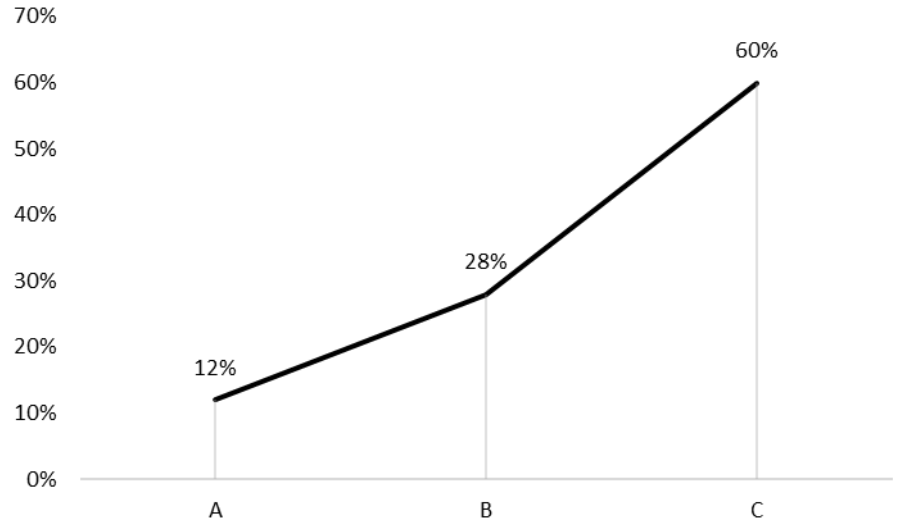

Fig. 1. Total percentages for incorrectly classified instances.

The majority of incorrectly classified instances from classes A and B, were classified by Random Forest in class C (60%). For classes A and C 28% were misclassified into B. We observe a low percentage, only 12%, of misclassifications from classes B and C to class A.

We note that Trans1 does not apply the basic syntactic rules, i.e. the subject-verb agreement, the subject-predicate agreement, as well as the modifiers agreements (attributive adjectives, predicate adjectives). Non-agreement is also observed in genitive constructions (possessive case, subjective and objective genitives), when of course there is not *of* (genitive case) or *by*. However, it has been found that Trans1 has, in many cases, a richer vocabulary than Trans2 and Trans3. In addition, Trans1 retains the main names, as Trans2 also does, in the Latin script, as it is considered right, and does the same in words not existing in its vocabulary, avoiding false and unrelated

translations, as in Trans2 and Trans3 sometimes occurs. Trans2 applies the above agreements, but not always successfully. Trans2 disposes quite satisfactory vocabulary, but not always about words that having more than two basic meanings. However, as it has been said, Trans2 translates all the common words, even those that do not exist in its vocabulary, breaking up compound words into their components and translating them, but, in some cases, this translation is wrong. Trans3 applies the above agreements more successfully than Trans2, it translates more successfully the components of compound words, but, as it has been said, Trans3 lags somewhat to the vocabulary richness.

It is important to know which features are more important to the classifier, so we tried the attribute evaluator technique (in Weka). Ratio of length in words and ratio of length in characters seem to be functional, as well as the Length Factor (LF), as we have described in section 2.3. On the other hand, comparisons, like if Ref=Trans1-2-3, seem not to be so useful for the classifier.

3 Conclusions and Future Work

In conclusion, this study aimed at automatic data selection for machine translation. It is based on the processing of a sufficiently large parallel corpora database. In this regard, we considered the data selection task as a classification problem. More specifically, three translation models were used, which represent both the old approach (SMT) and the state of art (NMT) to MT. In this way, differences in the translation process and the approach of the three models become more apparent. 82 characteristics have been calculated and 2,687 segments have been annotated. For proper analysis, we pre-processed our data before using Weka tool. We used Smote to address the class imbalance problem in our data. The results recorded give a better translational prediction to model 3, which does not make much of an impression, as this is a sophisticated translation model. It is worth mentioning that the translation was from English to Greek, which increased the task complexity, since the Greek language is a morohologically rich language with ambiguities. One way to more accurately approach ambiguities in the future might be the use of data categorization. For example, grammatical categorization may prove far superior to the lexical features employed herein, an approach that has already been considered for the Greek language [19]. Furthermore, it could be studied whether the use of in-depth features influences the translation process, such as etymology, that is believed to be of great help for the Greek language.

It is worth asking ourselves whether we can find similar results amongst other language pairs, and this may be a new field for study.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of 3th International Conference on Learning Representations, pp. 1-15. ICLR, San Diego (2015).

2. Barrón-Cedeño, A., Màrquez Villodre, L., Henríquez Quintana, C. A., Formiga Fanals, L., Romero Merino, E., & May, J.: Identifying useful human correction feedback from an online machine translation service. In: Proceedings of 23rd International Joint Conference on Artificial Intelligence, pp. 2057-2063. AAAI Press, Beijing (2013).
3. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 257–267. ACL, Austin (2016).
4. Breiman, L.: Random forests. Machine learning 45(1), 5-32 (2001).
5. Broder, A. Z.: On the resemblance and containment of documents. In: Proceedings of the Compression and Complexity of Sequences 1997, (pp. 21-29). IEEE Computer Society Washington, Washington (1997).
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321-357 (2002).
7. Daskalaki, S., Kopanas, I., & Avouris, N.: Evaluation of classifiers for an uneven class distribution problem. Applied artificial intelligence 20(5), 381-417 (2006).
8. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., & Dyer, C.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pp. 177-180. ACL, Prague (2007).
9. Koehn, P., Och, F. J., & Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48-54. ACL, Edmonton (2003).
10. Kordoni, V., Birch, L., Buliga, I., Cholakov, K., Egg, M., Gaspari, F., Georgakopoulou, Y., Gialama, M., Hendrickx, I.H.E., Jermol, M., Kermanidis, K., Moorkens, J., Orlic, D., Papadopoulos, M., Popovic, M., Sennrich, R., Sosoni, V., Tsoumakos, D., Van den Bosch, A., van Zaanen, M.; Way, A.: TraMOOC (Translation for Massive Open Online Courses): Providing Reliable MT for MOOCs. In: Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT), pp.376-400. European Association for Machine Translation (EAMT), Riga, (2016).
11. Miceli Barone, A. V., Haddow, B., Germann, U., Sennrich, R.: Regularization techniques for ne-tuning in neural machine translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1489- 1494. ACL, Copenhagen (2017).
12. Peris, Á., Cebrián, L., & Casacuberta, F.: Online Learning for Neural Machine Translation Post-editing. Cornell University Library arXiv preprint arXiv:1706.03196 1, 1-12 (2017).
13. Pighin, D., Màrquez, L. & May, J.: An Analysis (and an Annotated Corpus) of User Responses to Machine Translation Output. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, pp.1131-1136. European Language Resources Association (ELRA), Istanbul (2012).
14. Pouliquen, B., Steinberger, R., & Ignat, C.: Automatic identification of document translations in large multilingual document collections. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), pp.401-408. Recent Advances in Natural Language Processing (RANLP), Borovets (2003).
15. Sennrich, R., Firat, O., Cho, K., Birch-Mayne, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A., Mokry, J. & Nădejde, M.: Nematus: a toolkit for neural machine translation. In: Proceedings of the EACL 2017 Software Demonstrations, pp. 65–68. ACL, Valencia (2017).

16. Sennrich, R., Haddow, B., & Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1715-1725. ACL, Berlin (2016).
17. Sharaf, A., Feng, S., Nguyen, K., Brantley, K., & Daumé III, H.: The UMD Neural Machine Translation Systems at WMT17 Bandit Learning Task. In: Proceedings of the Conference on Machine Translation (WMT), pp. 667–673. ACL, Copenhagen (2017).
18. Singhal, S., & Jena, M.: A study on WEKA tool for data preprocessing, classification and clustering. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2(6), 250-253 (2013).
19. Stamatatos, E., Fakotakis, N., & Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational linguistics 26(4), 6-15 (2000).