How to Know the Best Machine Translation System in Advance before Translating a Sentence?

Bibekananda Kundu and Sanjay Kumar Choudhury
Language Technology, ICT and Services
Centre for Development of Advanced Computing, Kolkata
E-mail: {bibekananda.kundu,sanjay.choudhury}@cdac.in

Abstract

The aim of the paper is to identify a machine translation (MT) system from a set of multiple MT systems in *advance*, capable of producing most appropriate translation for a source sentence. *The prediction is done based on the analysis of a source sentence before translating it using these MT systems.* This selection procedure has been framed as a classification task. A machine learning based approach leveraging features extracting from analysis of a source sentence has been proposed here. *The main contribution of the paper is selection of source-side features.* These features help machine learning approaches to discriminate MT systems according to their translation quality though these approaches have no idea about working principle of these MT systems. The proposed approach is language independent and has shown promising result when applied on English-Bangla MT task.

1 Introduction

The question, *"Will machine translation ever replace human translation services?"*, is a long standing debate in the field of AI. Though state-of-the-art machine translation (MT) systems (Koehn, 2010; Koehn et al., 2007; Bach et al., 2007) often provide quite acceptable translations, but demand for producing high quality translations still requires professional human translators (HTs). In spite of debating on selection of slow, expensive but effective human translations versus fast and acceptable machine translations, collaboration of human and machine is a preferable option towards producing fast and high quality translations (of Edinburgh, 2014). In this collaboration, MT systems are used as aids for HTs to craft their translations. Several studies (Specia, 2011; Skadiņš et al., 2011; Koehn and Germann, 2014; Koponen, 2013) have provided strong evidence for improvement of productivity by post-editing machine translated outputs instead of unassisted translations. These aids enabling HTs to translate faster by simply adding, deleting words or reordering a small portion of the machine translation. As a result more and more high quality translations are being produced in shorter time. This synergy also speeds up the parallel corpora creation process (Chaudary et al., 2008) which eventually boosts the performance of statistical MT systems.

In addition to this, these days more than one MT systems are also available for certain language pairs. For instance, to translate from English to Bangla we have three MT systems, namely, AnglaBharati[1] (Sinha et al., 1995), Anuvadaksh[2] and Google Translator.[3] However, qualities of translations produced by individual systems are not the same. One possible reason behind this is they follow different paradigms. These paradigms in-

[1]Online system is available at http://tdil-dc.in/components/com_mtsystem/CommonUI/homeMT.php

[2]Description is available at http://tdil-dc.in/tdildcMain/IPR/Anuvaadaaksh.pdf

[3]Online system and description are available at https://translate.google.co.in/ and http://translate.google.co.in/about/intl/en_ALL/

clude rule-based (Bharati et al., 2010; Sinha et al., 1995; Magnúsdóttir, 1993), example-based (Carl et al., 2004), statistical (Koehn, 2010; Kunchukuttan et al., 2014) and hybrid (Sánchez-Cartagena et al., 2014; Chatterji et al., 2009). Moreover, in case of statistical MT systems, acceptability of translations differs according to the training set on which they are trained. Now an obvious question can come to our mind, *"How does one can use more than one MT system for translating from a source to target language?"*. Answer to this question is twofold: using consensus translation (Macherey and Och, 2007; Bangalore et al., 2001) and selecting most appropriate one from a set of translations generated by different systems (Dara et al., 2013; Zwarts and Dras, 2008). However, both approaches require same source sentence to be translated multiple times which indeed is a time consuming task. The time required to translate is proportional to the number of available MT systems. However, the processing time is not a matter of big concern because of availability of high speed computers and parallel computing algorithms. The main challenge is manual selection of suitable translation from a set of auto generated translations. To address this issue, we have proposed a methodology that automatically predict the best MT system in advance that will produce most appropriate translation for a source sentence. The prior prediction of best MT system does not require translation of the source sentence. It is also unaware of the working principle of the underlying MT systems. We hope that this approach will speed up the manual translation even more by reducing the searching cost for the appropriate translation.

1.1 What is the Value of Another Approach?

Solution addressing this specific problem is already in place. In this context, work of Sánchez-Martínez (2011) is worth mentioning. He has proposed a methodology to select the best MT system by using only information extracted from the source sentence to be translated without knowing the inner working of the MT systems. He has experimented on two European language pairs viz. English-Spanish and French-Spanish. He has used five binary maximum entropy classifiers and used mainly two types of features, namely, phrase-structure features and probabilistic features. The features reported by him, have also been used for sentence-level confidence estimation of MT (Quirk, 2004; Blatz et al., 2004).

1.1.1 Contributions

In this paper, we have proposed a similar approach to leverage the features extracted from source sentences without considering inner working principle of the MT systems. However, *our work is different from the existing approach with respect to selection of features and machine learning algorithms.* Therefore, two main ingredients of our proposed approach are (a) rich feature set and (b) classification technique. In addition to using phrase-structure and probabilistic features, *we have introduced dependency based features extracted from dependency trees.* Dependency based features have already been proven to be a good parameter for measuring goodness of machine translation (Bach et al., 2011; Goldberg and Orwant, 2013). *Moreover, to the best of our knowledge this is the first work done on English to Indian language MT systems.*

1.2 Outline

With a aim to select best MT system for a given source sentence, we have formulated this problem as a classification task in Section 2. Types of features used for this classification task have been discussed here. Visualization technique along with feature selection methodologies also have been discussed in this section. In Section 3, we explain the experiments conducted on two English-Bangla MT systems from two different paradigms: one is rule-based and the other one is statistical. Preparation of training dataset has also been discussed here. Results are discussed in Section 4. Finally, we conclude the paper with future scope of actions in Section 5.

2 Brief Grounding for Our Approach

The selection of best MT system from a multiple MT systems can be seen as a multi-class classification problem where class labels are $C = \{MT_1, MT_2, \ldots, MT_n\}$, representing different MT systems. Our task is to discrimi-

nate the best class $MT_i \in C$ using a machine learning classifier that depends on the features $F = \{f_1, f_2, \ldots, f_n\}$ extracted from analysis of source sentences. For simplicity, in this paper, we have framed this problem as a binary classification having only two class labels. In any classification task features play an important role to disseminate among classes. Therefore, in the next subsection we will discuss on the types of features used in this classification task.

2.1 Feature Set for Selecting Best MT System

Identification of proper features is a vital task to address this classification problem. In our case, the problem is quite difficult because we are trying identify features from the analysis of a source sentence to predict the quality of MT system. The main assumption is that source side features will be able to capture the latent relationships with the translation quality for that sentence. Source side features are extracted from phrase-structure and dependency trees. Figure 1 and 2 show a phrase-structure tree and a dependency tree of an English sentence *"Was my camera repaired already?"*. Features extracted from phrase-structure tree represent structural complexity of a sentence. Similarly features extracted from a dependency tree represent how words in a sentence depend on each other even for long distances. As a result, long distance dependencies are good indicators of complexity of a sentence. Based on training data, probabilistic features represent complexity in term of out-of-vocabulary (OOV), likelihood of a sentence, likelihood of a dependency relation, mapping capability of a source word to multiple target words or vice versa. Complexity related to phrase-structure and dependency influences the translation quality of rule-based MT systems whereas statistical MT systems are influenced by probabilistic features. So for proper classification, we have considered three types of features viz. (a) **phrase-structure features**, (b) **dependency features** and (c) **probabilistic features**. Some of the phrase-structure and probabilistic features have already been considered in the work of Sánchez-Martínez (2011). However, *we have introduced new features like number of unique Parts of Speech (POS) tags, POS tag density etc. in our phrase-structure feature set.* Examples of phrase-structure features are as follows:

- Number of Unique POS Tags (NUPT)
- POS Tag Density (PTD): PTD is a measure of the ratio of the number of different POS tags to the total number of POS tags.
- Maximum Depth of the Phrase-Structure Tree (MDST)
- Mean Depth of the Phrase-Structure Tree (MeanDST) : MeanDST is a measure of the ratio of the sum of the individual depths of all leaf nodes to the total number of leaf nodes.
- Maximum Number of Child Nodes per Node Found in the Phrase-Structure Tree (MNCNNFST)
- Mean Number of Child Nodes per Node Found in the Phrase-Structure Tree (MeanNCNNFST)
- Number of Internal Nodes (NIN)

Examples of probabilistic features are

- Joint Probability of Input Sentence (JPIS): We have approximated JPIS using trigram sequences as shown in Equation 1.

$$\begin{aligned} P(S = w_1 w_2 w_3 \cdots w_n) = & P(w_1) \\ & \times P(w_2|w_1) \\ & \times P(w_3|w_1 w_2) \\ & \times \cdots \\ & \times P(w_n|w_{n-2} w_{n-1}) \end{aligned} \quad (1)$$

- Joint Probability Using N-gram Dependency (JPUND): We have used dependency based language model as reported in (Shen et al., 2008). JPUND for the dependency tree shown in Figure 2 is cal-

culated using Equation 2.

$$\begin{aligned} JPUND = & P_T(repaired) \\ & \times P_L(camera \mid repaired_{head}) \\ & \times P_L(my \mid camera_{head}) \\ & \times P_L(was \mid my, camera_{head}) \\ & \times P_R(already \mid repaired_{head}) \\ & \times P_R(? \mid already, repaired_{head}) \end{aligned} \quad (2)$$

Here, $P_T(w)$ is the probability that word w is the root of a dependency tree. P_L and P_R are left and right side generative probabilities respectively.

- Maximum Fertility (MF) and Maximum Distortion (MD) : Fertility is defined as the number of target words generated from individual source words in a particular alignment of training data (Brown et al., 1990). Distortion is defined as the reordering distance of target words generated by a source word in a particular alignment (Brown et al., 1993). Depending on the target sentence alignment, each source word may have multiple fertilities and distortions. So maximum fertility and maximum distortion of a source sentence is defined as the highest fertility and distortion of individual constituent words of the source respectively. MF and MD of all vocabulary are pre-calculated during training. Later, these values are used to calculate the MF and MD of a source input sentence.

Features generated from dependency trees have already been used for measuring goodness of machine translation (Bach et al., 2011; Goldberg and Orwant, 2013). In this paper we have introduced dependency based features for selecting best translation system. Examples of some of our dependency based features are as follows:

- Number of Dependency Links (NDL)
- Maximum Dependency Distance (MDD): Dependency distance is the distance between head node and its dependent node. For example, in Figure 2 dependency distances between the head word *"repaired"* and its dependent words *"camera"*, *"already"* and *"?"* are 1, 1 and 2 respectively. MDD of the dependency tree shown in Figure 2 is 4 i.e. distance between *"ROOT"* and *"repaired"*.
- Mean Dependency Distance (MeanDD)
- Standard Deviation of Dependency Distances (SDDD)
- Maximum amongst the Number of Dependents of a Word (MNDW)

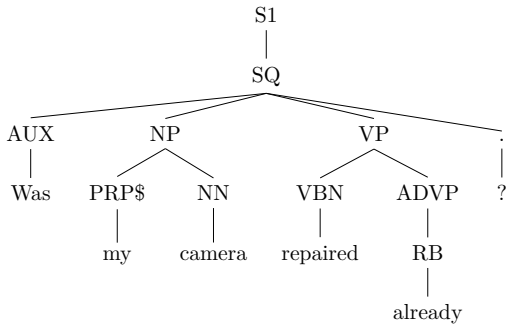

Figure 1: A phrase-structure tree.

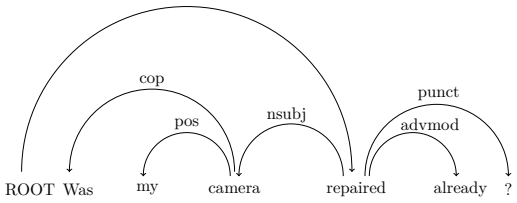

Figure 2: A dependency tree.

Now, using these three types of features we have extracted the feature values from the analysis of source sentences. An example of feature values extracted from the analysis of an English sentence, *"Was my camera repaired already?"*, has been shown in Table 1.

2.2 Feature Selection and Visualization

A large number of features have an adverse effect on efficiency and irrelevant features hamper the accurate prediction of class label. So, there is a requirement for reduction of dimensionality by filtering the irrelevant and redundant features. Manual identification of important features from a large number of features is practically not feasible. We have applied Information Gain (IG) (Lee and Lee, 2006)

Feature Attributes:	NDL	MDD	···	MNDW	NUPT	PTD	MDST	NIN	JPIS
Feature Values:	6	4	···	4	6	1.0	5	10	-14.4073

Table 1: An example of feature values extracted from the English sentence *"Was my camera repaired already?"*. Here, the joint probability value is shown in logarithm scale.

and Chi-square (χ^2) attribute evaluators (Jin et al., 2006) on our training data to find out features which are more relevant for this classification task. IG and χ^2 attribute evaluators evaluate the worth of an attribute by computing the value of the information grain and chi-squared statistic with respect to the class and rank features accordingly. IG and χ^2 attribute evaluators only provide rank of the features. Therefore, features having average merit value greater than a predefine threshold[4] are selected for the classification task. Impact of these features selected using different techniques, has been discussed in Subsection 3.2.3.

Visualization is important in the context of feature selection to visualize the underling representation and goodness of the training data (Feldman and Sanger, 2006). Visualization helps expert users to determine which features are important for their classification task. We have used a freely available visualization tool named as Hierarchical Clustering Explorer (HCE)[5]. Figure 3 shows mosaic view of the feature values in a representative sample of our training set[6]. In colour mosaic plot, graphical pattern of multidimensional data is represented using colourful tiles depending on the numerical value of data. High, low and medium values are represented using bright red, bright green and black colour respectively. As a value gets closer to the middle value from high to low or vice versa, the colour becomes darker. Here, each row represents individual feature values for all examples (training instances) whereas each column represents all feature values for individual instances.

[4]In our experiment we have manually selected the threshold values as 10±13 and 0.001±0.002 for χ^2 and IG attribute evaluators respectively

[5]HCE tool is available at `http://www.cs.umd.edu/hcil/hce/`

[6]Using HCE, interested readers can see the actual representation of the feature values in our training data which is available at `http://14.139.223.131/download/`

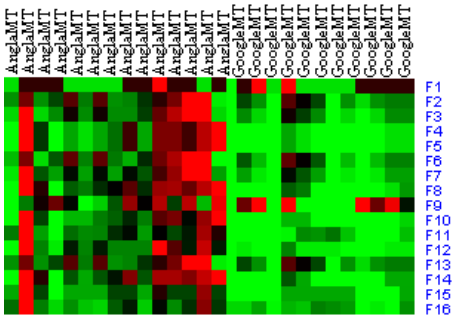

Figure 3: Visualization of multi dimensional features.

3 Experiments

3.1 Questions to Answer

Experiments in the current study are conducted to answer the following questions:

- Can features extracted from source sentences predict the quality of a MT system?
- Which machine learning algorithm is most appropriate for this classification task?
- How selection of different types of features influences the performances of classifiers?

In our work, the experiment has been conducted on English-Bangla MT systems.

3.2 Experiments on English-Bangla Machine Translation Systems

In our classification problem, we have two classes *"AnglaMT"* and *"GoogleMT"* representing AnglaBharati and Google MT system respectively. We have considered two MT systems belong to two different paradigms. The first one is a pseudo interlingua based (rule-based) MT approach and later one follows a statistical approach.

3.2.1 Data Preparation

We have collected 20k English sentences from Basic Travel Expression Corpus

(BTEC) (Takezawa et al., 2007) and manually translated them into Bangla. At the time of translating these sentences, two machine translations generated from AnglaBharati and Google MT systems have been provided to manual translators for their assistance. We also have collected a corpus of 50k English-Bangla parallel sentences which has been prepared for Health and Tourism domain under the Indian Languages Corpora Initiative (ILCI) project initiated by the DEITY, Govt. of India[7] (Jha, 2010). The first corpus contains sentences from basic expression dialogues of Travel domain. The second corpus contains 25k sentences from Health and 25k sentences from Tourism domain. Thus, we have collected total 70k English-Bangla parallel aligned sentences. Then we randomly selected 7k English sentences out of 70k and automatically translated them using both rule-based and statistical MT systems. For the rule-based MT we selected English to Bangla AnglaBharati system and for statistical MT we used Google MT system. Thus, we have prepared a dataset of 7k sentences having three translations of each English sentence i.e. one manual translation, two machine translations from a rule-based and a statistical MT system. Both the systems provide multiple translations. In case of multiple translations, we have selected the translation having highest BLEU (Papineni et al., 2002) score with the reference manual translation, so that the translation closest to the reference translation gets selected. After preparing these raw data having English sentences with their respective translations, we have applied all the feature functions on the English sentences to extract the feature values. Phrase-structure features (Zhang et al., 2008) are extracted from phrase-structure trees generated using Charniak parser (`http://cs.brown.edu/~ec/`). We have extracted dependency based features from dependency trees. Dependency trees are extracted using malt parser (`http://www.maltparser.org/`). Probabilistic features are extracted using Moses toolkit (Koehn et al., 2007) (`http://www.statmt.org/moses/`). To prepare the balanced dataset (Longadge and Dongre, 2013), we have selected 5906 instances[8] from 7k data, such that proportion of the classes become 50%. The final dataset has been prepared in *arff format* having 20 attributes, 2 classes and 5906 rows of instances. Each row contains 20 numerical feature values corresponding to 20 attributes. For the English sentence *"Was my camera repaired already?"*, an instance of the training data looks like $\{6, 4, \cdots, 4, 6, 1.0, 5, 10, -14.4073, AnglaMT\}$ (see Table 1). We assigned a class label *"AnglaMT"* if the BLEU score of the translation generated from AnglaBharati MT system is higher than translation generated from Google MT system otherwise, *"GoogleMT"* class label is assigned.

3.2.2 Experiment with Different Classifiers

We have used a WEKA data mining toolkit[9] (Witten and Frank, 2005) with default parameters for classifying MT systems based on the features extracted from the analysis of source sentences. Relevant features are extracted using an Information Gain Attribute Evaluator (IGAE) and a Chi-square Attribute Evaluator (CSAE) available in WEKA toolkit[10]. We have compared among different classifiers like WEKA implementation of Naïve Bayes, IBk (Aha et al., 1991) - a Lazy or Instance based learner that uses K-nearest neighbour algorithm, Multi Layer Perceptron (MLP) (Mitchell, 1997), LibSVM - a library for Support Vector Machines (Cristianini and Shawe-Taylor, 2000), Logistic - a class for building and using a multinomial logistic regression model with a ridge estimator (Le Cessie and Van Houwelingen, 1992) and Voted Perceptron (VP) (Freund and Schapire, 1999). These classifiers are trained on the dataset as discussed in Subsection 3.2.1. We have experimented using 10-fold cross validation on the same training

[7]The corpus is available at http://www.tdil-dc.in/

[8]Our dataset is available at `http://14.139.223.131/download/` so that researchers can experiment on it by applying their approaches.

[9]WEKA data mining toolkit is available at `http://www.cs.waikato.ac.nz/~ml/weka/`

[10]The class implementing IGAE and CSAE are `weka.attributeSelection.InfoGainAttributeEval` and `weka.attributeSelection.ChiSquaredAttributeEval`

data. To measure statistical significance of the performance of each of the classifiers, we have used the class implementing the Paired T-Tester (Mangal, 2012) which is `weka.experiment.PairedCorrectedTTester`.

3.2.3 Contributions of Feature Sets

We have designed our experiment to see the impact of individual types of features as well as combined effect on this classification task. We have conducted nine experiments using these WEKA implementations of classifiers on the same dataset. Each classifier is experimented using the following feature sets:

1. SF: a feature set contains only phrase-structure features.
2. DF: a feature set is prepared considering only dependency based features.
3. PF: a feature set having only probabilistic features.
4. SF + DF.
5. SF + PF: this is similar feature set used by Sánchez-Martínez (2011), except some extra phrase-structure features like NUPT, PTD etc.
6. DF + PF.
7. SF + DF + PF.
8. IGF: important features are extracted using IGAE.
9. χ^2F: relevant features are extracted using CSAE.

4 Results and Discussion

Figure 4 shows the performances of different classifiers using 10-fold cross validation and considering all feature values. These performances are measured in term of Percentage Correct. It has been seen that the nearest neighbour classifier i.e. IB1 outperforms any other classifier considered. It uses normalized Euclidean distance to find the training instance closest to the given test instance and predicts the same class according to this training instance. We have measured the contributions of different feature sets in terms of Percentage Correct obtained using 10-fold cross validation of the same training data. The results of our experiments on different feature sets have been shown in Table 2. Standard deviations of Percentage Corrects of each of the classifiers are shown within brackets. Moreover, these values are statistically significant within 95% confidence intervals. From the experimental results, we can see that the performance constantly improves when we used combined feature sets. Moreover, performances of all the classifiers except IBk improve for the feature type SF+DF+PF. However, best performance has been achieved using IB1 for the feature type SF+PF. Though the introduction of DF does not substantially improve the performance of IB1, but it has been observed that this introduction helps to improve the performance of other classifiers. For instance, we can see from Table 2, an overall increment of 2.5 for LibSVM, 0.97 for MLP and 1.05 for VP has been achieved.

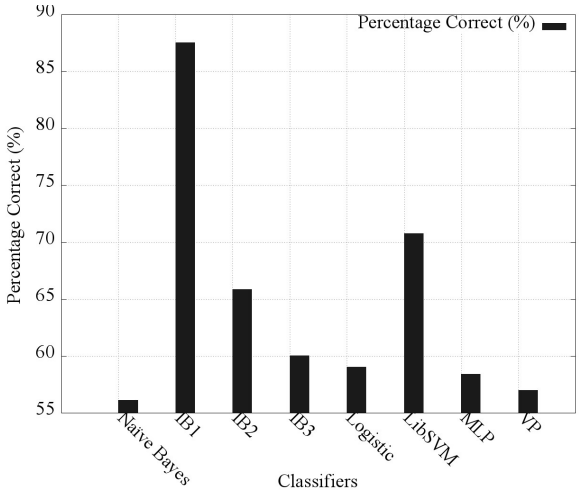

Figure 4: Comparison among performances of different classifiers for choosing best MT system.

5 Conclusions

In this paper, we have proposed a machine learning approach for selecting a MT system producing most appropriate translation before translating the input sentence. The prediction is done depending only on the features extracted from a source sentence without knowing the inner process of the MT system. Our approach uses phrase-structure, probabilistic and dependency features. Primarily, the proposed approached has been ap-

Feature Set	Classifiers							
	Naïve Bayes	IB1	IB2	IB3	Logistic	LibSVM	MLP	VP
SF	56.09 (1.97)	84.34 (1.56)	66.31 (1.54)	60.59 (1.74)	57.72 (2.22)	60.56 (1.86)	56.56 (2.75)	55.49 (2.06)
DF	56.00 (1.96)	80.10 (1.69)	66.87 (1.90)	62.42 (1.96)	57.08 (2.01)	59.18 (1.71)	54.95 (2.56)	55.75 (1.86)
PF	55.70 (1.88)	87.31 (1.30)	65.73 (1.52)	59.78 (1.71)	56.55 (2.32)	61.55 (2.10)	55.15 (2.70)	53.11 (1.96)
SF+DF	**56.23** **(1.98)**	86.14 (1.70)	66.20 (2.07)	60.19 (2.05)	58.27 (2.00)	66.44 (2.07)	57.69 (2.47)	57.13 (1.91)
SF+PF	55.99 (1.92)	**87.77** **(1.36)**	66.60 (1.56)	60.51 (1.87)	**59.09** **(2.22)**	68.28 (1.81)	57.43 (2.31)	55.93 (1.88)
SF+DF+PF	56.14 (1.98)	**87.55** **(1.38)**	65.85 (1.84)	60.05 (2.02)	**59.02** **(2.56)**	**70.78** **(2.10)**	**58.40** **(2.56)**	56.98 (1.93)
IGF	55.78 (1.94)	87.03 (1.47)	65.11 (1.82)	59.82 (2.08)	58.86 (2.26)	69.86 (2.07)	57.64 (2.64)	**57.64** **(2.64)**
χ^2F	55.83 (1.95)	87.11 (1.38)	65.07 (1.93)	60.51 (2.01)	58.73 (2.27)	69.71 (1.89)	57.80 (2.42)	56.49 (1.92)

Table 2: Contributions of different feature sets measure in Percentage Correct (%).

plied on English-Bangla MT systems. Experiment shows IB1 classifier provides statistically significant performance (with 95% confidence) considering all types of features. The proposed approach is language independent and can be applied on any language pair where multiple MT systems are available. Features used in this paper can also be applied on similar NLP tasks where measuring confidence of the system is required.

5.1 Where from Here?

We think the next step should be examining the human and machine translations in parallel to have insight into which features are more central in determining quality of MT systems. Given these features, we would like to employ an ensemble classifier (Dietterich, 2000) to combine the predictions of multiple classifiers. Because ensembles can often perform better than any single classifier. As a future work, we are also planing to use a development set to tune the trained classifiers. Moreover, we would also like to use different configuration of WEKA to see the changes in performances of individual classifiers. In this classification task, we have seen that the best performance is achieved by the IB1 which is basically an instance based classifier. The competence of instance based learners depends on several factors viz. size, position of instances in multidimensional feature space and inherent problem complexity due to decision boundary (Massie, 2006). Therefore, we are also planing to concentrate on these factors for further improvement of the performance of this classifier.

References

David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-Based Learning Algorithms. *Machine Learning*, 6(1):37–66, January.

Nguyen Bach, Matthias Eck, Paisarn Charoenpornsawat, Thilo Khler, Sebastian Stker, ThuyLinh Nguyen, Roger Hsiao, Alex Waibel, Stephan Vogel, Tanja Schultz, and Alan Black. 2007. The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System. In *Proc. of the International Workshop on Spoken Language Translation*, Trento, Italy.

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A Method for Measuring Machine Translation Confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 211–219, Stroudsburg, PA, USA.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing Consensus Translation from Multiple Machine Translation Systems. In *Automatic Speech Recognition and Un-*

derstanding, 2001. ASRU '01. IEEE Workshop on, pages 351–354.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 2010. *Natural Language Processing: A Paninian Perspective*. PHI.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Comput. Linguist.*, 16(2):79–85, June.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguist.*, 19(2):263–311, June.

Michael Carl, Andy Way, and Walter Daelemans. 2004. Recent Advances in Example-Based Machine Translation. *Comput. Linguist.*, 30(4):516–520, December.

Sanjay Chatterji, Devshri Roy, Sudeshna Sarkar, and Anupam Basu. 2009. A Hybrid Approach for Bengali to Hindi Machine Translation. In *Proceedings of 7th International Conference on Natural Language Processing (ICON-2009)*, pages 83–91, India.

Sriram Chaudary, Kiran Pala, Lakshminarayana Kodavali, and Keshav Singhal. 2008. Enhancing Effectiveness of Sentence Alignment in Parallel Corpora: UsingMT & Heuristics. In *Proceedings of International Conference on Natural Language Processing (ICON-2008)*, India. Macmillan Publishers.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-base Learning Methods*. Cambridge University Press.

Aswarth Abhilash Dara, Sandipan Dandapat, Declan Groves, and Josef van Genabith. 2013. TMTprime: A Recommender System for MT and TM Integration. In *HLT-NAACL'13*, pages 10–13.

Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15, London, UK. Springer-Verlag.

Ronen Feldman and James Sanger. 2006. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA.

Yoav Freund and Robert E. Schapire. 1999. Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, 37(3):277–296, December.

Yoav Goldberg and Jon Orwant. 2013. A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *Second Joint Conference on Lexical and Computational Semantics, Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA, June.

Girish Nath Jha. 2010. The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may.

Xin Jin, Anbang Xu, Rongfang Bie, and Ping Guo. 2006. Machine Learning Techniques and Chi-square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. In *Proceedings of the 2006 International Conference on Data Mining for Biomedical Applications*, BioDM'06, pages 106–115, Berlin, Heidelberg. Springer-Verlag.

Philipp Koehn and Ulrich Germann. 2014. The Impact of Machine Translation Quality on Human Post-editing. In *Workshop on Humans and Computer-assisted Translation*, pages 38—46, Gothenburg, Sweden, April.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, USA, 1st edition.

Maarit Koponen. 2013. This Translation Is Not Too Bad: An Analysis of Post-editor Choices in a Machine-Translation Post-editing Task. In *Proceedings of Workshop on Post-editing Technology and Practice*, pages 1–9.

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-Anuvadak: Tackling Multiway Translation of Indian Languages. In

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014).

Saskia Le Cessie and JC Van Houwelingen. 1992. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201.

Changki Lee and Gary Geunbae Lee. 2006. Information Gain and Divergence-based Feature Selection for Machine Learning-based Text Categorization. *Information Processing and Management*, 42(1):155–165, January.

Rushi Longadge and Snehalata Dongre. 2013. Class Imbalance Problem in Data Mining Review. *CoRR*, abs/1305.1707.

Wolfgang Macherey and Franz J. Och. 2007. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. In *Proceedings of the 2007 Joint Conference on EMNLP-CoNLL*, pages 986–995, 209 N. Eighth Street, East Stroudsburg, PA, USA.

Guorñ Magnúsdóttir. 1993. Review of "An Introduction to Machine Translation" by W. John Hutchins and Harold L. Somers. Academic Press 1992. *Computational Linguistics*, 19(2):383–384, jun.

S. K. Mangal. 2012. *Statistic in Psychology and Education*. PHI, India.

Stewart Massie. 2006. *Complexity Modelling for Case Knowledge Maintenance in Case-based Reasoning*. Ph.D. thesis, The Robert Gordon University, December.

Tom Mitchell. 1997. *Machine Learning*.

University of Edinburgh. 2014. SMT Research Survey Wiki: a comprehensive survey of statistical machine translation research publications. `http://www.statmt.org/survey/Topic/ComputerAidedTranslation`. [Online; accessed 7-August-2014].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, pages 311–318, Stroudsburg, PA, USA.

Christopher B. Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 825–828, May.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2014. The UA-Prompsit Hybrid Machine Translation System for the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 178–185, Baltimore, Maryland, USA, June.

Felipe Sánchez-Martínez. 2011. Choosing the Best Machine Translation System to Translate a Sentence by Using only Source-Language Information. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 97–104, Leuve, Belgium. European Association for Machine Translation.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *Proceedings of Association for Computational Linguistics*, pages 577–585.

RMK Sinha, K. Sivaraman, Aditi Agrawal, Renu Jain, Rakesh Srivastava, and Ajai Jain. 1995. ANGLABHARTI: a Multilingual Machine Aided Translation Project on Translation from English to Indian Languages. In *Systems, Man and Cybernetics Intelligent Systems for the 21st Century., IEEE International Conference*, volume 2, pages 1609–1614, October.

Raivis Skadiņš, Maris Puriņš, Inguna Skadiņa, and Andrejs Vasiljevs. 2011. Evaluation of SMT in Localization to Under-resourced Inflected Language. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011)*, pages 35–40, Leuven, Belgium.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *15th Conference of the European Association for Machine Translation*, EAMT, pages 73–80, Leuven, Belgium.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12:303–324, September.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Min Zhang, GuoDong Zhou, and Aiti Aw. 2008. Exploring Syntactic Structured Features over Parse Trees for Relation Extraction Using Kernel Methods. *Information Processing and Management*, 44(2):687–701, march.

Simon Zwarts and Mark Dras. 2008. Choosing the Right Translation: A Syntactically Informed Classification Approach. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1153–1160.