Journal of Advances in Science and Technology
Vol. 17, Issue No. 1, March-2020, ISSN 2230-9659

Data Sparsity and Unsupervised Training of the Classifier-Based Speech Translation

Durga Gupta[1]*, Dr. Vijay Singh[2]

[1] Research Scholar, Shri Krishna University, Chhatarpur M.P.

[2] Associate Professor, Shri Krishna University, Chhatarpur M.P.

Abstract - We focused on speech recognition jobs that need huge volumes of tagged voice data yet are challenging to gather. Both academic study and practical applications make use of domain-adaptive and semi-supervised learning techniques. Algorithms from the field of machine learning may be utilised for unsupervised learning, which entails studying and categorising data sets that have not been labelled. It is hypothesized that the accuracy of the suggested strategy depends on the size of the n-best lists. The trials employed n-best lists with sizes of 100, 500, 1000, and 2000 to make these findings. Thus, these "unsupervised" algorithms are able to find patterns in data without any external supervision. This study made use of a dataset that was compiled throughout the duration of the Transonics project. In addition, the output vocabulary may benefit greatly from employing a more robust SMT engine. For this purpose, we have adopted a strategy for determining how far apart two statements are conceptually, and a suitable clustering algorithm. To deal with the problem of sparse data, researchers have developed a novel approach that use statistical machine translation software to train classifiers.

Keywords - Data, Sparsity, Unsupervised, Speech, Translation

----------X----------

1. INTRODUCTION

It is widely accepted that phrase-based Statistical Machine Translation (SMT) techniques are the primary translation methodology to use when developing S2S translation systems. To function effectively, they rely on a local mapping of potentially lengthy word sequences. These approaches' adaptability allows them to handle a wide range of situations in the dialogue space. However, accuracy in translation cannot be guaranteed. This is exacerbated when the input is a "noisy" utterance riddled with inflection and other speech defects. Unfortunately, the input quality of an SMT system is typically severely degraded due to lexical and syntax problems. Everything is interrelated and affects the accuracy with which concepts are translated and the naturalness of the speech synthesised in the target language. Because of this, it is standard practice to use SMT approaches in tandem with other translation strategies. The major goal of interactive S2S applications is not to provide a word-for-word (literal) translation of the source, but rather to facilitate natural conversation between the interlocutors. Most current SMT tools prioritize literal translation over such interpretation. When the source utterances are categorized according to the ideas they communicate, the work of translation is reduced to assigning a sample phrase in the target language for each concept class.

Classifying ideas is analogous to quantifying a continuous "semantic" sub-space. It is safe to assume that the classifier is doing its job if the quantization error is minor and the input utterance is inside the quantizer's coverage domain, and the selected idea is an excellent fit. It is obvious that it is not possible to completely classify the area of conversation. The reason for this is because there is an endless number of possible ideas to be discussed during a conversation, which would need a very high quantization level. There will be a decline in classifier accuracy if there are a great many idea classes to choose from. It follows that the classifier has to be paired with a considerably more versatile translation system, such as an SMT engine. One way to determine whether an input utterance is beyond the classifier's coverage is to use a rejection mechanism. Traditional pattern recognition applications, such as voice and picture recognition, have seen widespread adoption and use of deep neural networks (DNN). In the beginning, researchers concentrated on acoustic models constructed using deep neural networks and the hidden Markov model (DNN-HMM). Recently, research has shifted its attention to end-to-end speech recognition systems, which do away with HMM entirely in favor of DNN. Numerous speech-recognition products on the market now use it. Higher performance in DNN-based acoustic models is dependent on a large quantity of training data and

the inclusion of more parameters than in traditional HMM-based models.

2. LITERATURE REVIEW

Garcia Martinez, et al (2020) Current machine translation systems still struggle with translating across morphologically rich languages. To combat the data sparsity issue brought on by factors such as data availability (quantity), domain shift, and the languages at play, this research explores a number of different neural machine translation (NMT) architectures (Arabic and French). We demonstrate that even with a vast amount of data, Using linguistically motivated components, like in the Factored NMT model, can achieve better results than traditional NMT systems based on subword units by more than 1 BLEU point. When applied to NMT, our study demonstrates the value of linguistic characteristics in both low- and high-resource settings.

Li, H., Liu, Y., Mamoulis, N., & Rosenblum, D. S. (2019) One of the primary functions of recommender systems is sequential suggestion, which involves predicting the user's future action based on the user's previous activities. Recently, a translation-based sequential recommendation technique was developed using the Translating Embeddings approach to knowledge graph completion. When dealing with difficult translations, we find that TransRec's ability to provide correct recommendations is impaired. Keeping this in mind, we offer a translation-based recommender that uses temporally dynamic relaxation and category-specific projection to cater to customers' varying preferences.

Sharma, Manoj & Chaudhary, Nalin & Khubchandani, Sagar (2018) We provide an overview of computational theory in this work. Our conversation has progressed to a crucial stage in the theory of computing. To put it simply, the field of study known as "Theory of Computation" draws from both the realms of mathematics and computer science. Computing hardware and software are where the theory of computing is most often applied. Complexity Theory, Computability Theory, and Automata Theory are the three subfields that make up the Theory of Computation.

Riding, Jon & Boulton, Neil (2017) Extremely huge training data sets are required for each of these systems. Since most Bible translation is done in languages with little resources, we lack this information. While there is often very little or no bilingual corpus available at the onset of a project, this article presents an innovative approach to employing computers as Machine Assisted Translation (MAT) engines, which help the translator from the very beginning of a project. This necessitates the development of systems that can start learning from very little inputs and build up their knowledge base until it can sustain more conventional MT procedures. A potential framework for doing this is described, and preliminary experiment results are explored.

Yogi, Kuldeep & Jha, Chandra & Dixit, Shivangi (2015) Except for the most basic of translations, the results of machine translation are not accurate enough for direct usage. Not every word is translated literally, but the overall sense of a statement is conveyed. Given the growing importance of MT, it's important to be able to evaluate the accuracy of MT results. Many leading experts in MT-Research are working to develop an MT-System capable of producing high-quality translations in a wide variety of target languages. To a large degree, time and money may be saved if only the best translations are used. In the future, only high-quality translations will be passed on to post-editing, while the remainder will undergo either pre-editing or a retranslation. In this article, the likelihood of machine translation output is determined using the Kneser Ney smoothing language model. However, determining a translation's accuracy is impossible. Many other elements than the probability rating are important in determining a hypothesis's overall quality. Using two distinct specified renowned algorithms for categorization, the quality of machine translation may be more accurately estimated during post-editing.

3. RESEARCH METHODOLOGY

- **Data Used**

Experiments were conducted to evaluate the suggested strategy and its efficacy was compared to supervised training. The current dataset was initially compiled for use in the Transonics project. The following steps were taken to compile the data set: Expert opinions and medical dictionaries were used to choose the first ideas. The goal of the Transonics project was to develop a speech-to-speech translator from English to Farsi for use in clinical settings, and this study makes use of data acquired for that purpose. To accommodate the medical staff's needs, 1,269 distinct idea classes were hand-picked with input from specialists and medical dictionaries. The technique was tested using just the 97 most heavily paraphrased English classrooms, out of a total of 1,269. After that, approximately half (500) of the source utterances (in English) were used for training, whereas the other half (707) were used for testing. To create n-best lists in the target language, our team used a phrase-based SMT that was educated on a parallel English/Farsi dataset including 1.2M English words. Constructing n-best lists with 500, 1000, 2000, and 3000 hypotheses per seed phrase allowed us to examine how these different list sizes might affect our results.

One thousand phrases were chosen at random from the aforementioned collection and used as the test corpus for this study. When picking words for the test set, we excluded one per class to guarantee that all classes were adequately represented in the

training set. The SMT was trained on a Farsi-English parallel dataset consisting of 148K lines in English. Using this information, we also created bilingual classification background models. The SMT was fine-tuned with the aid of a parallel development set that had 915 lines from the English side.

The g-means software was used to implement the spherical k-means algorithm as the foundation of the clustering process. To achieve this goal, vector models were first generated from the documents using the MC toolkit (n-best lists). In addition, In order to evaluate our approach, we compared it to one in which the 97 output groups were assigned utterances at random(random clustering). The outcomes of these tests using various clustering techniques are shown in Table 1. The Exchange strategy was tested using both the KLD and JSD measures. The SMT n-best list length was set to 2,000 for these tests. Table 1 also includes findings from the spherical k-means clustering technique. Both the original English sentences and the n-best list texts were analysed using K-means for comparison. In Table 1, Results from random clustering and supervised training with data annotations are shown side by side. The levels of agreement in clustering and the average entropy were measured across all cases.

Table 1: The results of different clustering schemes

Method	Agreement [%]	1-Agreement [%]	Ave. Entropy [bits]	Acc. [%]	Acc. 4-best [%]
Random	97.5	2.5 (baseline)	3.785	14.3	33.9
Exchange method with KLD (n = 2, 000)	98.4	1.6 (36%)	5.158	45.7	63.6
Exchange method with JSD (n = 2, 000)	98.4	1.6 (36%)	5.115	47.8	61.7
Spherical k-means with original data	97.9	2.1 (16%)	4.763	38.3	52.1
Spherical k-means with n-best documents	98.1	1.9 (24%)	4.843	37.3	53.7
Reference annotation	100.0	0	6.213	69.4	86.0

Since a classifier-based translator was the primary objective, each of the aforementioned scenarios culminated in the training of a classifier using the generated clusters and the subsequent evaluation of its correctness using testing data. Besides only gauging how accurate the classifier was overall, we also looked at how accurate it was inside its top four outputs (Table 1).

Table 2: SMT n-best list using Exchange method with JSD metric

n-best length	500	1,000	2,000	3,000
Agreement [%]	98.3	98.3	98.4	98.3
Ave. Entropy [bits]	5.060	5.091	5.115	5.088
Accuracy [%]	44.1	45.0	47.8	46.0
Accuracy within 4-best [%]	61.2	60.4	61.7	60.1

Experiments with varying n-best list sizes are summarized in Table 2. With the Exchange method employing the JSD metric, we were able to get the average entropy and the clustering agreement for each individual test. Table 2 displays results for the accuracy and accuracy within 4-best outputs of the tested classifiers.

- **Experiments**

Classifiers based on both the suggested technique and the standard method were evaluated side-by-side. It is hypothesized that the accuracy of the suggested strategy depends on the size of the n-best lists. The trials employed n-best lists with sizes of 100, 500, 1000, and 2000 to make these findings. Table 1 displays the findings. In all of these tests, the background interpolation factor was set at 0.9, which is quite near to the optimal number. The influence of the interpolation factor was investigated over a range of values in both the traditional and novel methods. For the classic method, 500 candidates were chosen as the best possible list size. Adjustments in precision as a function of the interpolation factor for these two approaches are shown in Figure 1.

- **Clustering evaluation**

In this study, we used two distinct strategies for gauging the efficacy of the clustering procedure. This means that for every pair of items in the data, depending on whether or not the reference supports the premise that they are part of the same cluster, the outcome might be true or false. The average entropy of clusters, a metric of their purity, is the basis of the second way of assessment. The average entropy of cluster set C is defined as, where R is the collection of reference classes.

$$E = -\sum_{c \in C} \frac{|c|}{|C|} \sum_{r \in R} P_{cr} \log(P_{cr})$$

Where,

$$P_{cr} \triangleq \frac{|c \cap r|}{\left| c \cap \left(\bigcup_{\rho \in R} \rho \right) \right|}$$

4. DATA ANALYSIS

Several results are shown in Table 1 that may be utilised to better create an unsupervised training method. The fifth column indicates that the classification accuracy for the suggested technique is greater than that for the spherical k-means method. Training under supervision and without it is night and day. Columns 2–4 of Table 1 provide the clustering assessment metrics. In the third column, we see that although 1Agreement is a useful relative statistic, the agreement measure itself has become useless owing to its quick saturation. Conversely, average entropy fluctuation follows the same pattern as that seen in classification accuracy. Therefore, It appears useful as a yardstick for reaching one's objectives.

There was further evidence that 4-best classifier outputs were around 2% more accurate when the KLD measure was used instead of the JSD metric. Since a result, the former measure is preferable for apps that provide the user many choices, as this improves the likelihood of selecting the correct one. Some ideas are significantly more common than others, hence the utterances do not have a uniform distribution throughout the classes. Both the training and test data sets retain this characteristic, which means certain categories will have more items than others. These things are so common that they would dominate certain clusters even in a haphazard arrangement. While being tested, they aid in the proper categorization of things into their most common groups, with an accuracy of 14.3 percent (random clustering). In Table 2, we can see how the n-best list's size affects the classifier's performance. Clustering quality increased from 500 to 2000 for n-best list length, but dropped to 1000 for n-best list length of 3000. Adding more SMT hypotheses to the clustering process increases lexical variety, leading to better outcomes. At some point, however, it becomes false to assume that all of the possibilities are accurate renditions of the original remark. Hypotheses of the lowest quality towards the bottom of the list bring about the worst clustering results. The measures of clustering assessment shown in Table 2 an demonstrate a similar trend, but at a weaker level of significance.

According to Table 3, the proposed method may decrease the relative error rate by as much as 10.4 percentage points. But as predicted, this number falls as SMT n-best lists become longer because of the increasing noise in lower-ranked SMT outputs. Table 3 displays the proportion of valid 4-best classifier outputs for each approach. In that case, the error rate of the recommended approach was 23.7% lower than that of the gold standard approach. When the SMT n-best list reached 1,000 items long, it was the highest level of accuracy within 4-best. Furthermore, as the n-best list became longer, the accuracy dropped because the classifier models had become noisier. In Figure 1 we can see how the background model influences the classifier's performance. In the picture, when comparing the recommended and default methods, There is a relationship between the background interpolation factor, the accuracy of the 4-best outputs, and the accuracy of the one-best output (). Since all class scores are determined by the background model, the curves show that the classifier is unable to differentiate between classes when =0. A little tweak to, however, yields a huge boost in precision.

This is because the background model was developed using a wide, general-purpose corpus, thus it does not discriminate based on class. When is small λ, the background model's score is more important than any other in determining the final classification. Despite this, the class LMs have a considerable impact on the scores and hence the classifier's performance. The importance of the class LMs λ grows as rises. Figure 1 demonstrates how this improves the accuracy and discriminative power of classifier models. In situations when the factor is close to one, classes with spiky models and low vocabulary coverage emerge as a consequence of the background model's smoothing effect being reduced. When equals one, accuracy λ rapidly declines. Figure 1 displays that the suggested approach and the traditional method both follow the aforementioned pattern, but the new method continues to be better over the whole range of λ that was investigated. At λ= 0.999, the greatest measured accuracies for both the traditional and new approaches were 75.2% and 78.7%, respectively. Thus, the suggested technique had a significantly (14.1%) reduced error rate than the standard method.

Table 3: Comparing the performance of the traditional approach with the new strategy in terms of classification accuracy

	Conventional (baseline)	n-best length			
		100	500	1000	2000
Accuracy [%]	74.9	77.4	77.5	76.8	76.4
Relative error reduction [%]	0.0	10.0	10.4	7.6	6.0
Accuracy in 4-best [%]	88.6	90.7	91.0	91.3	90.5
Relative error reduction [%]	0.0	18.4	21.1	23.7	16.7

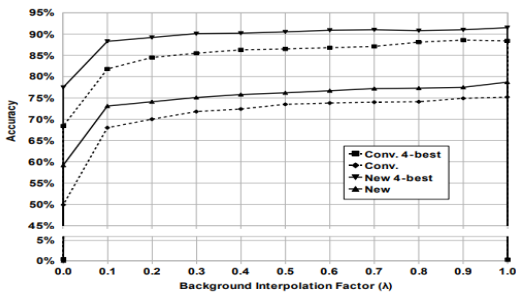

Figure 1: Implications of the Context Model on Classification Precision

Figure 1 shows that when accuracy is evaluated relative to the four best outputs, the proposed

method outperforms the standard method. At the λ= 0.9 sample point, For the traditional method, 88.6 percent was found to be the greatest possible accuracy for the top four predictions. At the λ= 0.999 sample point, the suggested approach attained a value of 91.5%. If we compare the suggested method's error rate to that of the 4-best classifiers, we find that it is 25.4% lower.

5. CONCLUSION

The suite of speech-to-speech translation software has benefited greatly from the usage of concept categorization as a translation strategy. However, It takes a lot of time to annotate large amounts of training data, which is the main barrier to such classifiers achieving good results. Because it might be challenging to gather a big quantity of labeled voice data for certain speech recognition tasks, we focused on them. Both academic study and practical applications make use of domain-adaptive and semi-supervised learning techniques. The suggested technique allows for the automated training of the idea classifier. Unfortunately, the results are far less accurate than those obtained during supervised training. For classifier-based translation systems, this study represents a crucial first step toward the development of unsupervised training approaches. Research is underway to find ways to improve the training process, such as by semi-automatically evaluating the class sizes and applying filtering at different points. Experimental results showed that it outperformed a baseline classifier that had not been exposed to the original source language paraphrases. Any time the input sentence is properly classified, the classifier acts as a filter to smooth out the SMT output and provide natural-sounding translations. We also think that employing a more feature-rich SMT engine might provide substantial improvements by expanding the range of words available in the final product.

REFERENCES

1. Garcia Martinez, Mercedes & Aransa, Walid & Bougares, Fethi & Barrault, Loïc. (2020). Addressing Data Sparsity for Neural Machine Translation Between Morphologically Rich Languages. Machine Translation. 34. 10.1007/s10590-019-09242-9.

2. Li, H., Liu, Y., Mamoulis, N., & Rosenblum, D. S. (2019). Translation-Based Sequential Recommendation for Complex Users on Sparse Data. IEEE Transactions on Knowledge and Data Engineering, 1–1. doi:10.1109/tkde.2019.2906180

3. Sharma, Manoj & Chaudhary, Nalin & Khubchandani, Sagar. (2018). An Introduction of Theory of Computation. 10.23883/ijrter.2018.4156.axkvi.

4. Riding, Jon & Boulton, Neil. (2017). Learning from Sparse Data.

5. Yogi, Kuldeep & Jha, Chandra & Dixit, Shivangi. (2015). Classification of Machine Translation Outputs Using NB Classifier and SVM for Post-Editing. Machine Learning and Applications: An International Journal. 2. 21-29. 10.5121/mlaij.2015.2403.

6. Ettelaie, E., Georgiou, P.G., Narayanan, S., 2010. Hierarchical classification for speech-to-speech translation. In: Proc. of the Interspeech, Makuhari, Japan.

7. Papineni, K., Roukos, S., Ward, T., Zhu, W., 2012. Bleu: a method for automatic evaluation of machine translation, Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

8. Abdel-Hamid, Ossama, Deng, L., Yu, D., Jiang, Hui, 2013. Deep segmental neural networks for speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. pp. 1849–1853.

9. Ali Ahmed, Vogel Stephan, Renals Steve Speech recognition challenge in the wild: Arabic MGB-32017 IEEE Automatic Speech Recognition and Understanding, Workshop, ASRU, IEEE (2017), pp. 316-322

10. Aradilla Guillermo, Vepa Jithendra, Bourlard H ervé, Using Posterior-Based Features in Template Matching for Speech Recognition: Technical Report IDIAP (2006)

Corresponding Author

Durga Gupta*

Research Scholar, Shri Krishna University, Chhatarpur M.P