Linguistically Informed Transformers for Text to American Sign Language Translation

Abhishek Bharadwaj Varanasi, Manjira Sinha, Tirthankar Dasgupta, Charudatta Jadhav
TCS Research
[1]{varanasi.abhishek,sinha.manjira,dasgupta.tirthankar,charudatta.jadhav}@tcs.com

Abstract

In this paper we propose a framework for automatic translation of English text to American Sign Language (ASL) which leverages a linguistically informed transformer model to translate English sentences into ASL gloss sequences. These glosses are then associated with respective ASL videos, effectively representing English text in ASL. To facilitate experimentation, we create an English-ASL parallel dataset on banking domain. Our preliminary results demonstrated that the linguistically informed transformer model achieves a 97.83% ROUGE-L score for text-to-gloss translation on the ASLG-PC12 dataset. Furthermore, fine-tuning the transformer model on the banking domain dataset yields an 89.47% ROUGE-L score when fine-tuned on ASLG-PC12 + banking domain dataset. These results demonstrate the effectiveness of the linguistically informed model for both general and domain-specific translations. To facilitate parallel dataset generation in banking-domain, we choose ASL despite having limited benchmarks and data corpus compared to some of the other sign languages.

1 Introduction

Sign Languages (SL) are the primary means of communications for the deaf community. It is a non-verbal form of communication where deaf individuals use their hands, arms, and facial expressions to share thoughts and ideas. Unlike spoken languages that rely on sound, ASL employs gestures. Recent linguistic studies have confirmed that SLs, like other spoken languages, is a complete natural language with its own syntactical structures and intricate morphological and phonological properties. This complexity includes both sequential and simultaneous affixation of manual and non-manual elements in its structure.

A challenging aspect of sign language translation (SLT) is that Sign Languages (SLs) are multi-

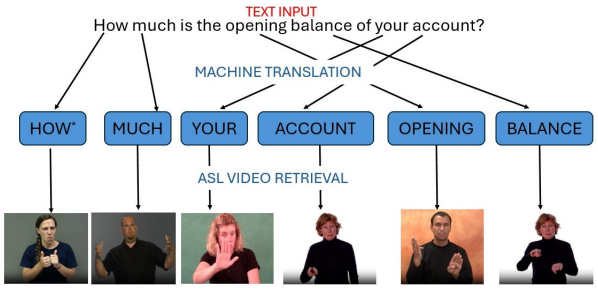

Figure 1: Illustration of text to American Sign Language (ASL) translation using glosses as intermediate step.

channeled and do not have a written form, as noted by (Langer et al., 2014). Consequently, advancements in text-based machine translation (MT) cannot be directly applied to SLs. Historically, researchers have used written representations of SLs to facilitate translation. One common method involves using glosses, which are labels in the spoken language that correspond to sign language components, often including affixes and markers. These glosses act as an intermediary step in developing MT systems that translate between SLs and spoken languages (noted by (Cihan Camgoz et al., 2017; Camgoz et al., 2018); (Chen et al., 2022)), and vice versa ((Stoll et al., 2020); (Saunders et al., 2020)). Other notable works include (Stoll et al., 2020) that proposed an approach using Neural Machine Translation (NMT) and motion graphs to generate sign language videos for a given text. (Moryossef et al., 2023) have proposed a method of converting the text to sign language glosses, extracting the poses for each gloss and translating the poses to a video. In an earlier attempt, (Dasgupta and Basu, 2008) have proposed a method translating text to Indian Sign Language (ISL) using Lexical Functional Grammar while (Sugandhi et al., 2020) talks about generating animated avatar using *HamNoSys* for text to SL translation. For translations from spoken languages to SLs, glosses are used to build the system in two phases: translating text

to glosses, and then converting these glosses into video (see Figure 1). These glosses are then input into systems that generate SL content, such as avatar animations or auto-encoder based video generators. Our present research specifically targets the text-to-gloss translation phase, which is crucial for producing accurate sign language animations. However, despite improvements in this area, significant breakthroughs remain elusive, as indicated by Rastgoo et al. (2021).

In this paper we propose a framework for automatic translation of English text to ASL. The key contributions and results of this work are as follows:

1. We leverage a novel method called linguistically informed transformer architecture that takes into account both the word level and different linguistic feature embeddings using a Graph Convolution Network (GCN) for the MT task. The primary focus is to translate English sentences into ASL gloss sequences. These glosses are then associated with respective ASL videos, effectively representing English content in ASL.
2. To facilitate experimentation, we have manually created an English-ASL parallel dataset on banking domain. Banks play a pivotal role in the daily lives of individuals impacting personal finance and economic stability. Hence, facilitating communication for the deaf community in the banking domain is essential. The dataset will be released with this paper.
3. Our preliminary results demonstrated that the linguistically informed transformer model achieves a 97.83% ROUGE-L score for text-to-gloss translation on the ASLG-PC12 dataset. Furthermore, fine-tuning the transformer model on the banking domain dataset yields an 89.47% ROUGE-L score when fine-tuned on ASLG-PC12 + banking domain dataset.

The above results show a significant improvement from the baseline GRU-B model. ASLG-PC12 (Othman and Jemni, 2012), the largest text-to-ASL gloss dataset, offers 87,710 samples but pales in comparison to mainstream language pairs like English to French. Its focus on news and politics limits its applicability across domains, necessitating domain-specific models, increasing data scarcity challenges.

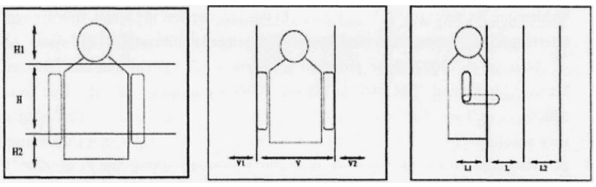

Figure 2: Classification of signing space into horizontal, vertical, and lateral regions.

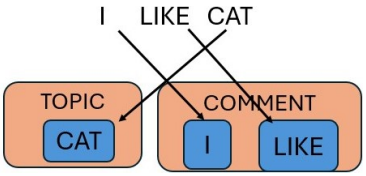

Figure 3: Illustration of Topic-Comment Structure

2 Sign Language (SL) Linguistic Issues

Sign Languages (SL) are visual-spatial natural languages, utilizing manual and non-manual components for linguistic communication (Zeshan, 2003). Manual components include hand shape, orientation, position, and movement, while non-manual components consist of facial expressions, eye gaze, and body posture. Signers utilize a three-dimensional signing space segmented into 27 cubical regions (Sinha, 2003, 2009). Each sign formation adheres to complex constraints akin to spoken languages, with SL morphology primarily derivational. The closed lexical class in SL encompasses classifier hand shapes, discourse markers, and non-manual signs (Zeshan, 2003). Classifier hand shapes offer specific hand configurations representing referent characteristics, as shown in Figure 2.

American Sign Language (ASL) is known to follow a topic-comment structure. This structure positions the main subject or theme (the topic) at the sentence's outset, followed by more specific information (the comment) (Struxness, 2010). By establishing context early in the sentence, ASL users efficiently convey complex ideas (Figure 3). One important aspect of ASL's topic-comment structure is the flexibility in word ordering. While the default order is topic-comment, ASL allows for variations based on emphasis and context. For instance, a speaker might choose to emphasize a particular aspect of the comment by placing it before the topic. This flexibility adds nuance and richness to ASL communication, enabling speakers to convey subtle meanings and emotions effectively (Struxness, 2010). The flexibility in the structure is the reason

why a simple rule-based approach is not possible for text to ASL gloss translation.

3 The Linguistically Informed Transformer for Text to ASL Gloss

The existing NMT models excel at capturing intricate data patterns without requiring manual feature engineering, offering end-to-end solutions. However, they often overlook latent linguistic traits crucial for extracting pertinent information. To address this, we propose a transformer-based architecture that integrates word embeddings from the encoder part with diverse linguistic features inherent in text, enhancing automatic text-to-ASL gloss translation.

3.1 Transformer Model

The input to the model is a sentence consisting of a word sequence $x = (x_1, x_2, ..., x_T)$ representations. We then tokenize the sentence x using a wordpiece vocabulary, and then generate the input sequence $\bar{x}$ by concatenating a [CLS] token, the tokenized sentence, and a [SEP] token. Then for each token $\bar{x}_i \in \bar{x}$, we convert it into vector space by summing the token, segment, and position embeddings, thus yielding the input embeddings $h^0 \in R^{(n+2)\times h}$, where h is the hidden size. Next, we use a series of L stacked Transformer blocks to project the input embeddings into a sequence of contextual vectors $h^i \in R^{(n+2)\times h}$. Here, we omit an exhaustive description of the block architecture and refer readers to Vaswani et al. (2017) for more details.

3.2 Syntactic Dependency Graph

Encoding the structural information directly into neural network architecture is not trivial. Marcheggiani and Titov (Marcheggiani and Titov, 2017) proposed a way to incorporate structural information into sequential neural networks through Graph Convolution Networks (GCN) (Webster et al., 2019; Kipf and Welling, 2016). GCNs take graphs as inputs and conduct convolution on each node over their local graph neighborhoods. The syntax structure of a sentence is transferred into a syntactic dependency graph, and GCN is used to encode this graph information. This kind of architecture is already utilized to incorporate syntactic structure with BERT (Devlin et al., 2018) embeddings for several NLP based tasks (Duvenaud et al., 2015).

3.3 Linguistically Informed Transformer

We have incorporated the similar method for the present text-gloss translation task in this work. Here, each sentence is parsed into its syntactic dependencies graph and use GCN to consume this structural information. We use pre-trained GLOVE embeddings as our initial hidden states of vertices in GCN. The output hidden states of the GCN is combined with the context embeddings generated by the transformer model's (T5 and BART) encoder and then passed to the decoder unit.

4 Experiments

Dataset: The ASLG-PC12 corpus (Othman and Jemni, 2012): consists of 87,710 bilingual sentences. It contains 1,027,100 English words and 906,477 gloss words, along with 4,662 English word singletons and 6,561 gloss word singletons. The vocabulary for both sign gloss annotation and spoken language comprises 16,788 and 12,344 terms, respectively.

ASL-Bank Dataset: Considering the specificity of the terminology used in banking contexts, we have also built a set of 3597 text- ASL gloss pairs through domain experts. The collected phrases are sourced from banking-related texts and provided to American Sign Language (ASL) experts for manual translation into ASL gloss. Refer Appendix D for data statistics and Appendix E for sample data.

Fine-tuning: We divided the ASLG-PC12 corpus into 52,626 sentences for training, 17,542 sentences for validation, and 17,542 sentences for testing (Amin et al., 2021) and use it to fine-tune a T5-small, T5-base and BART-base model on A100 GPU for 50 epochs (experiment A). For the ASL-Bank dataset, we used 3,166 sentences for training, 395 sentences for validation and 396 sentences for testing. We fine-tuned the three transformer models from experiment A on A100 GPU for 50 epochs. Refer Appendix A and B for more details.

Evaluation: Apart from using the standard MT evaluation parameters like, ROUGE-L (Lin, 2004) and BLUE (Papineni et al., 2002) scores we also advocate using a modified BERTScore (Zhang et al., 2019) as performance metrics. As the BERT models are trained on natural English text, we cannot rely on the sentence embeddings it gives for the ASL gloss sequences for the reasons explained in introduction. Hence, we proposed to get the word embeddings of each gloss present in the ASL gloss sequence and aggregate them to get the sentence

embedding of the ASL gloss sequence which can be further used to calculate the cosine similarity score.

5 Results

The results are reported by first comparing the model performance upon fine-tuning on ASLG-PC12 (Othman and Jemni, 2012) dataset between our models of choice T5-small, T5-base (Raffel et al., 2020) and BART-base (Lewis et al., 2019) and a GRU based model from Amin et al. (2021) (Table 1).

Model	GRU-B (Amin et al., 2021)	T5-small*	T5-base*	BART-base*
ROUGE-L	94.37	97.82	**97.83**	97.36
BLEU-1	93.26	97.68	**97.73**	97.21
BLEU-2	89.64	97.16	**97.22**	96.51
BLEU-3	86.68	96.36	**96.43**	95.53
BLEU-4	83.98	94.65	**94.76**	94.66
***Modified* BERTScore**	—	**97.59**	97.58	97.55

Table 1: Comparing scores on ASLG-PC12 test dataset for text to gloss with other work

From table 1, it is clear that the transformer models BART-base, T5-small and T5-base are better performing compared to a GRU model. Further, we check how well these fine-tuned transformer models are performing on our ASL-banking dataset Since ASLG-PC12 dataset has no samples related

Model	T5-small*	T5-base*	BART-base*
ROUGE-L	59.06	**59.13**	57.71
BLEU-1	60.66	**60.98**	59.93
BLEU-2	37.44	**37.74**	36.64
BLEU-3	26.18	**26.42**	25.48
BLEU-4	19.66	**19.8**5	19.01
Modified BERTScore	87.45	**87.64**	87.37

Table 2: Test scores when tested on our ASL-banking dataset using the T5-small*, T5-base* and BART-base* models

to banking domain, the scores drop when tested on our banking dataset (Table 2). Hence, we have further fine-tuned the BART-base model, T5-base model and T5-small model on our ASL-banking dataset (Table 3). We also checked if including an equal number of samples from ASLG datatset (i.e., 3597 samples) along with our ASL-banking dataset improves the test scores and we observed that there is a significant improvement in the test scores (Table 4).

T5-base model is the best performing transformer model for text-to-ASL translation task on

Model	T5-small	T5-base	BART-base
ROUGE-L	78.25	**79.96**	77.92
BLEU-1	81.11	**83.08**	79.80
BLEU-2	64.45	**67.89**	64.19
BLEU-3	53.86	**58.14**	54.14
BLEU-4	46.69	**51.47**	47.27
Modified BERTScore	92.06	**92.16**	92.13

Table 3: Comparing scores on our ASL-Banking test dataset for text to gloss using transformer models after further fine-tuning

Model	T5-small	T5-base	BART-base
ROUGE-L	89.17	**89.47**	85.50
BLEU-1	90.67	**90.83**	86.83
BLEU-2	82.48	**83.0**5	77.41
BLEU-3	76.87	**77.74**	70.67
BLEU-4	72.47	**73.58**	65.50
Modified BERTScore	**93.38**	93.37	93.09

Table 4: Comparing scores on our ASL-Banking test dataset using the transformer models fine-tuned on ASLG + ASL-Banking dataset

both ASLG-PC12 dataset and our ASL-banking dataset. A few challenges in the text to gloss translation task are: In some cases, word ordering is different within the topic part and the comment part of the predicted texts compared to the gold texts. In case of *wh*-questions, the *wh* word is sometimes placed at the beginning of the sentences and sometimes at the end. In a few sentences, helping verbs and articles are not removed. So, they should be exclusively removed using SpaCy's parts-of-speech tagging. Few words are being replaced by their synonymous words in the gloss translations. It's not a problem while signing the text, but it is reducing the scores of metrics like ROUGE-L and BLEU.

6 Conclusion

In this paper, we present a linguistically informed transformer architecture towards automatic translation of English text to American Sign Language. The proposed model not only aims at addressing the poor generalization capability of traditional structured prediction models but also exploit the linguistic characteristics present within a text to improve the performance of the translation. We evaluate the performance of the proposed model with respect to a popular baseline model from Amin et al. (2021). We observed that the proposed transformer based model along with an additional linguistic information performs much better than existing baseline system.

Limitations

1. This work specifically focuses on text to ASL gloss translation. Hence, the fine-tuned models cannot be used for generating glosses in other sign languages like Indian sign language or British sign language due to differences in structure.
2. As shown in Figure.1, the glosses generated using the proposed framework can be mapped to videos (refer Appendix C) and can be streamed together. But the output has inconsistencies due variation in resolution and people signing from video-to-video. This can be tackled with video generation which is not in the scope of this work.
3. Since the syntactical structure of sign language is very much different from that of natural language, open-source LLMs like LLaMA family can be leveraged by combining external sign language rules. It also helps in tackling the limitation of using a single model for different sign language translations. This can be achieved with techniques like Retrieval Augmented Generation (RAG) but this is not in the scope of this work.

References

Mohamed Amin, Hesahm Hefny, and Mohammed Ammar. 2021. Sign language gloss translation using deep learning models. *International Journal of Advanced Computer Science and Applications*, 12(11).

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3056–3065.

Tirthankar Dasgupta and Anupam Basu. 2008. Prototype machine translation system from text-to-indian sign language. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 313–316.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Gabriele Langer, Susanne König, and Silke Matthes. 2014. Compiling a basic vocabulary for german sign language (dgs)–lexicographic issues with a focus on word senses. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, pages 767–786.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. An open-source gloss-based baseline for spoken to signed language translation. *arXiv preprint arXiv:2305.17714*.

Achraf Othman and Mohamed Jemni. 2012. English-asl gloss parallel corpus 2012: Aslg-pc12. In *Sign-lang@ LREC 2012*, pages 151–154. European Language Resources Association (ELRA).

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 of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Samar Sinha. 2003. A skeletal grammar of indian sign language. *Unpublished master's diss., Jawaharlal Nehru University, New Delhi, India*.

Samar Sinha. 2009. *A grammar of Indian sign language*. Ph.D. thesis, PhD dissertation, Jawaharlal Nehru University, New Delhi, India.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Kevin Struxness. 2010. American sign language grammar rules. Technical report, Technical Report,[on-line: http://daphne. palomar. edu/kstruxness/Spring ....

Sugandhi, Parteek Kumar, and Sanmeet Kaur. 2020. Sign language generation system based on indian sign language grammar. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–26.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kellie Webster, Marta R Costa-Jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (gap) shared task at the gender bias in nlp workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7.

Ulrike Zeshan. 2003. Indo-pakistani sign language grammar: a typological outline. *Sign Language Studies*, pages 157–212.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Hyperparameters used for fine-tuning on ASLG-PC12 dataset

1. Training epochs: 50
2. Learning rate: 1e-5
3. Weight Decay: 1e-6
4. Warm-up Epochs: 10
5. Batch size: 10
6. Gradient accumulation steps: 4
7. Optimizer: Adam

B Hyperparameters used for fine-tuning on our ASL-Banking dataset

1. Training epochs: 50
2. Learning rate: 1e-5
3. Weight Decay: 1e-5
4. Warm-up Epochs: 10
5. Batch size: 2
6. Gradient accumulation steps: 4
7. Attention Dropout: 0.1
8. Optimizer: Adam

The linguistic embeddings which are GCN's output hidden states are combined with the last hidden state of the encoder part as described in section 3.3 during both the fine-tuning processes. In that way, the GCN is trained along with the transformer model.

C Video Retriever

The generated ASL gloss sequence is tokenized into individual glosses using SpaCy tokenizer and each gloss is mapped to its corresponding ASL video which is stored in a folder/database. If there is no video match for a particular gloss, we check if it is a common noun or an adjective. If yes, then we try to find a video for its synonym. We use NLTK word-net to find the synonyms. The synonyms are sorted in lexicological order. We iterate through this list and check if there exists a video each of synonym. As soon as we find a synonym which has a video, we break the loop and use this video for signing the original word. If it is neither a common noun nor an adjective or there is no video even for any of its synonyms, we will simply sign it letter-by-letter (as shown in Figure 4).

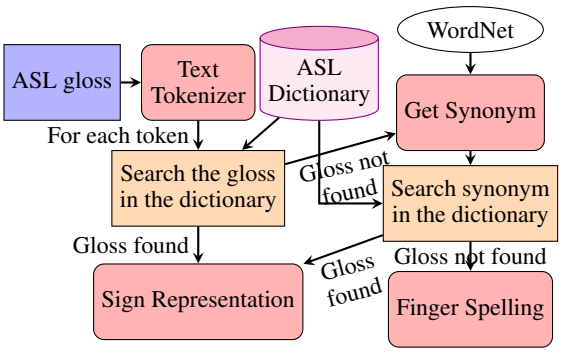

Figure 4: Video Retrieval Process flow

D Data Statistics

Conversational Type	Vocabulary Size	Type	Count
Banker point-of-view	772	Declarative	446
		Interrogative	168
Customer point-of-view	1595	Declarative	502
		Interrogative	2488
		Total:	3597

E Sample Data

English Text	ASL Gloss Sequence
A basic savings account it is. I'll need you to fill out this form with your personal details.	SAVINGS ACCOUNT BASIC, IT IX. FORM THIS FILL-OUT NEED YOU, YOUR PERSONAL DETAILS IX-loc.
A basic savings account it is. I'll need you to fill out this form with your personal details.	SAVINGS ACCOUNT BASIC, IT IX. FORM THIS FILL-OUT NEED YOU, YOUR PERSONAL DETAILS IX-loc.
Can I deposit a check via mobile banking?	DEPOSIT CHECK MOBILE BANKING CAN I?
How can I assist you in updating your contact information for your account?	HOW CAN I ASSIST YOU IN UPDATING YOUR CONTACT INFORMATION FOR YOUR ACCOUNT ?
Your credit check came back clear, so we can proceed with finalizing your account.	CREDIT CHECK YOUR FINISH, CLEAR. ACCOUNT YOUR FINALIZE CAN PROCEED WE.