==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 20:8 August 2020
==

Generating a Parallel Corpus Stream for Odia: Mining Parallel Corpus from Odia Twitter

Anjan Kumar Panda, MSC IT, KSOU Mysore
Internet Application Specialist, Technology Manager
Life Member, OSA. The Odisha Society of the Americas
5050, Hacienda Drive, Apt 2232, Dublin, CA, 94568
panda.anjankumar@gmail.com
Contact: 1- 845-535-0961

Dr Arun Kumar Malik, PhD, Assistant Professor of Political Science
Gujarat National Law University, Gandhinagar
amalik@gnlu.ac.in
Contact No. 8128650850

Note: This research article is part of a sequence of papers to enable the researchers in the field of computational linguistics in Odia.

Introduction

A corpus is a fundamental need for natural language process applications.

A parallel corpus is a foundational need for languages like Odia (Oriya - The Unicode Standard, Version 13.0."https://unicode.org/charts/PDF/U0B00.pdf. Accessed 8 Aug. 2020) which would enable explorations in natural language processing advancements into machine translation, (Machine translation - Wikipedia. "https://en.wikipedia.org/wiki/Machine translation. Accessed 8 Aug. 2020) computational language modelling, (Language model - Wikipedia."

https://en.wikipedia.org/wiki/Language_model. Accessed 8 Aug. 2020) Question Answer Systems, Generative (Better Language Models and Their Implications - OpenAI." 14 Feb. 2019, https://openai.com/blog/better-language-models/. Accessed 8 Aug. 2020) systems.

To make a neural model learn actively a stream of training data is needed.

NLP tasks based on neural architectures based on deep learning need a lot of training data.

Machine Translation tasks need millions of parallel pairs known as a parallel corpus for training.

This paper describes a way to mine a parallel corpus stream on social media to be used by machine learning-based natural language processing systems.

Methodology

Why parallel corpus

A large Parallel corpus (Parallel corpora - Ilc-Cnr." http://www.ilc.cnr.it/EAGLES96/corpustyp/node20.html. Accessed 8 Aug. 2020) is one of the input components needed by Natural language processing tasks like Machine Translation. "Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data" (Natural language processing - Wikipedia." https://en.wikipedia.org/wiki/Natural_language_processing. Accessed 8 Aug. 2020.). "Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another" (Machine translation - Wikipedia." https://en.wikipedia.org/wiki/Machine_translation. Accessed 8 Aug. 2020)

"Parallel corpora are made in the business of communication in multilingual societies," Present SOTA Machine Translation models (Machine Translation, Papers with Code." https://paperswithcode.com/task/machine-translation. Accessed 8 Aug. 2020) uses deep learning (Deep learning - Wikipedia." https://en.wikipedia.org/wiki/Deep_learning. Accessed 8 Aug. 2020) and neural network (Artificial neural network - Wikipedia." https://en.wikipedia.org/wiki/Artificial_neural_network. Accessed 8 Aug. 2020) architectures to predict a translation from an input. These models using deep learning need a lot of training data, in case of a machine translation engine, the training data comes from parallel corpora.

India is a multilingual society. Present advances in technology and the fields of linguistics allows Indian language systems to advance towards easing the multilingual communication channels. Amongst Indian languages, absence of large parallel corpus acts as a blocker for the research in Odia computational linguistics.

How to make a pair and a corpus

To make a parallel pair, an input in one language gets translated to output in another language retaining the original meaning.

"I love you" in English can be translated to "ମୁଁ ତୁମକୁ ଭଲପାଏ"

An example of a parallel pair is followed

- "I love you": "ମୁଁ ତୁମକୁ ଭଲପାଏ".

A collection of related pair is called corpus, here is one with eight parallel pairs.

- "I love you": "ମୁଁ ତୁମକୁ ଭଲପାଏ"
- "ତୁମେ ମୋ ପ୍ରିତୀ ର ଶ୍ରେୟ" : "I love You"
- "ତୁମେ ମୋ ପ୍ରଣୟ ଆଶା" : "I love You"
- "ମୁ ତୁମ ପ୍ରଣୟ ଭୀକ୍ଷୂ" : "I love You"
- "I am an Odia" : "ମୁଁ ଜଣେ ଓଡ଼ିଆ"
- "I am an Odia" : "ମୁଁ ଓଡ଼ିଆ"
- "ମୁଁ ଓଡ଼ିଆ କହେ": "I speak Odia"
- "ମୁଁ ଓଡ଼ିଆ କହିପାରେ" : "I can speak Odia"

The usefulness of these pairs

New Machine translation systems can be trained on corpora described above.

Benchmarks like “language modelling benchmark” (Language Modelling, Papers with Code." https://paperswithcode.com/task/language-modelling. Accessed 8 Aug. 2020.) are used all over the applied linguistics (Applied linguistics - Wikipedia." https://en.wikipedia.org/wiki/Applied_linguistics. Accessed 8 Aug. 2020) research space. We believe a reasonable sized parallel corpus can scale natural language processing research for Odia. Researches on comparative linguistics can use this corpus for understanding the language comparisons.

While “Google Translate” has recently launched Odia into its supported languages, (Google Translate supports new languages for the ... - The Verge." 26 Feb. 2020, https://www.theverge.com/2020/2/26/21154417/google-translate-new-languages-support-odia-tatar-turkmen-uyghur-kinyarwanda. Accessed 8 Aug. 2020) other big technology organizations are yet to support Odia. Google and others are way behind when implementing Odia.

We believe new generations of researchers and applications may use this corpus as a base to build their research. So in our view, a readily available parallel corpus is useful as a fundamental building block of language research.

Why parallel corpus Stream?

Effectiveness and accuracy of the language models depend significantly on the size of a large corpus that is used as training data. “The Effects of Corpus Size and Homogeneity on Language Model Quality” (The Effects of Corpus Size and Homogeneity on Language ...." https://www.aclweb.org/anthology/W97-0118.pdf. Accessed 8 Aug. 2020) is a well-researched topic described in this paper for speech recognition systems. Language tasks like Machine Translation need a lot of quality corpus for exactly the same reason. In the age of Big Data and Deep Learning, A big corpus gives better quality of a model. Deep learning models use the big corpus to learn features from the first few layers of the network and then learn to predict tasks in deeper layers before generating output. There is a question for many languages like Odia, which are underrepresented computationally at present in the digital world with a small if available corpus.

Where will we get a large parallel corpus?

A stream is a solution in this situation.

A stream is a solution to building a large corpus, incrementally, while continuing the research possible with a small corpus. For example, a synthesizer network can be trained on a small corpus to generate synthetic corpus that in-turn can be used as a training pair for a translation network.

A stream will also be an effective input mechanism for the neural networks that learn actively, they learn from new training data continuously.

A large corpus will be an accumulation of a continuously generating stream of pairs into it. Many large corpora can be created by specialists if we can learn and model to generate one corpus by the stream.

Why social media is an effective medium for streaming parallel pairs

New user-generated content is available in Social media continuously. The Facebook posts, twitter feeds are essentially data streams generated from users, travel through many paths on the internet and ultimately stored in servers as data stores.

Due to these properties of current, user-generated, continuous nature of the data generated on social media, it is chosen as the catchment area for the stream of the parallel corpus.

Why twitter is chosen for the stream

Twitter is chosen because of its unique characteristics as noted below.

It provides a small entry point (240 character limit), avenues of engagement (reply, poll), programming interfaces (APIs).

With this limitation of character, a parallel pair can be in the form of a tweet, Odia and English part divided in it. A parallel pair can also be derived from a tweet where a translation sub quotes the original tweet.

These two approaches to derive a parallel pair can be encoded in the tools like regular expressions available in various computer languages

The tweet Style:

- <Beginning><English Sentence><separator><Translated Odia Sentence><End>
- <Beginning><Odia Sentence><separator><Translated English Sentence><End>
- <Beginning> <Translation Tweet> <Sub Tweet><End>

Twitter gives Application Programing Interfaces to programmatically parse the tweet streams, extract the pairs using the above-mentioned style guide.

Why not another form of social media like Facebook

While character limitations on twitter was a primary motivator for choosing it as the first catchment social media platform to mine parallel corpus streams, other social media platforms like Facebook can potentially be used.

To act as a stream catchment following properties are required.

- Continuous content generated from User Base
- Programmability
    - APIs to read the user content
    - Style Guide to parse the parallel pairs

Procedure: Steps to extract parallel pairs and suggested usage

1. **Origination**: a pair can be generated in a few ways
    a. An Odia and English knowing Twitter user tags a parallel pair in a tweet
    b. A volunteer twitter account creates a parallel pair in a tweet
2. **Streaming**: All tweets having parallel pair form a stream having sequential nature
3. **Mining**: Corpus is mined from the stream by
    a. A tweet bot software to read the generated stream
    b. A software module inside the bot to extract parallel pairs from the stream
    c. A software module inside the bot to store the extracted pairs into a storage
4. **Validation**
    a. The pairs are community validated during the origination

   b. Before feeding pairs to the machine translation models a community of academicians to validate

5. **Usage**
   a. The corpus stream may be used by NLP engines by consuming the generated pairs in JSON format sampled below, or CSV format by applying data transformation algorithms.

      {en: “I love you”:

      or: “ମୁଁ ତୁମକୁ ଭଲପାଏ”}

Why this is effective

This model is effective because

- It involves a Natively Odia Speaking Community to generate content
- It involves a Natively Odia Speaking moderator community to moderate and validate content
- It involves Academics to the validated generated stream
- It uses technical community to build the tools required for the complete pipeline
- It creates a foundation for the research community

This model's effectiveness can be measured from the result

So far the following is observed from a bot on twitter named as ଶୁଆ (Sua) (ଶୁଆ (@mte2o) | Twitter." https://twitter.com/mte2o. Accessed 8 Aug. 2020) that follows the model we described.

- This model is now operational as a tweet bot.
- It generates around 100 parallel pairs per day.
- Till 15th May 2018, it has generated ~12800 English Odia pairs.

Problem formulation

Need is to generate as many parallel pairs as possible, in a continuous stream format.

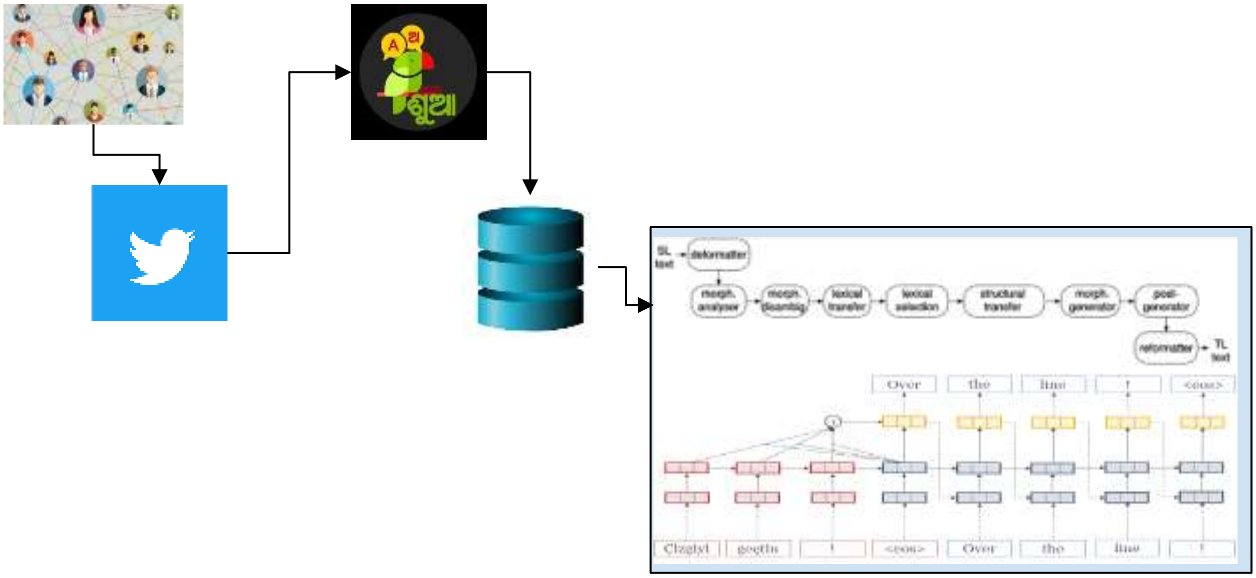
Twitter Interface provides a small entry point. "The text content of a Tweeter can contain up to 240 characters (Counting characters — Twitter Developers." https://developer.twitter.com/en/docs/basics/counting-characters. Accessed 8 Aug. 2020), avenues of engagement (reply, poll), programming interfaces (APIs) It can go to the methodological part and step ways description required like how you extract pair words, justification of extracting from social media.

To generate a parallel corpus, in our experience, using social media is one of the most effective approaches due to its properties of engagement, ease of use, usage levels, and availability of multiple formats to use them. Social media reflects contemporary subjects by user's contributions from their own fields of experience and most often reflects what is happening at the moment, it includes diversity such as art from artists, public policy from public administrators in charge or a short story and poetry from writers, users skilled at various levels from a novice to an expert in their field participate in social media. Social media allows everyone to participate at their skill level. If searched deeper, social media has historical artefacts to get historical artefacts too. While it is not a great place for quality academic research, it is a great place to socialize and have digital content generated in the form of conversations in natural languages. Hence, we believe its effectiveness being used as a catchment platform for generating streams of the parallel corpus.

Out of many Social media platforms twitter is chosen as it is a buzzing place, everyone contributes to that space, the tweets like the replies. The characteristics of the platform and its suitability as the first catchment platform for corpus stream generation is listed above in "why twitter is chosen".

Odia subset of twitter is full of participants who know both English and Odia, understand what a parallel pair means and can easily tune themselves to contribute pairs. For example, there are accounts like @OdiaCulture ("Odia Culture (@OdiaCulture) | Twitter." https://twitter.com/odiaculture. Accessed 8 Aug. 2020) who tweets on Odia culture @CMO_Odisha (CMO Odisha (@CMO_Odisha) | Twitter." https://twitter.com/cmo_odisha. Accessed 8 Aug. 2020) who tweets on Odisha. There are many international accounts like billionaire space and electricity entrepreneur @elonmusk whose tweets can be translated by the Odia community to make meaningful pairs. Though these reference accounts are used initially, eventually a community builds around an account and it starts to take its own personality.

Community contributions of the pairs generate a continuous stream of pairs sequenced on time which can be consumed by actively learning agents built on neural network based architectures as presented indicatively in the following diagram.



Hence getting a parallel stream of parallel pairs of corpus is a function of the process of the engagement with Odia Twitterati, process of picking the candidate parallel pairs, the algorithms and computational toolset to extract the pairs and convert it into a stream.

1. Community
    a. Community tweets on their own
    b. Some of them tweet bilingual pairs
    c. Translating Community translates some tweets from community
2. Tweeter: All these tweets goes to twitter infrastructure
3. ଶୁଆ Tweet Bot: Parses the tweet stream and generates parallel corpus stream
4. NLP systems: read the stream generated by the tweet bot

Consolidating the discussion above may we present the notation and the formulation?

- 2 : Parallel Corpus Stream =PaCoSt
- 1: Community Odia twitterati = CoOrT

- 3: Each eligible pair = Ep
- 3: Algorithms =algo
- 3: 4,5 : ToolSet = ts

The Corpus Stream Equation

PaCoSt=process (CoOrT, Ep, algo, ts)

May we call these adobe as the corpus stream equation? Parallel corpus stream is a function of the process involving Community, contributing eligible parallel pairs, presented as a stream, parsed by tools like tweetbots and regular expressions involving algorithms to read stream, parse and extract pairs and store into a data store or pass the stream to a subscribing actively learning neural network .

Solution

Community Odia Twitterati (*CoOrT)*

Odia Community in twitter was observed to be having following properties to be chosen as a community base to generate a parallel corpus stream

- Participating members
  - Anyone knowing both the languages, English and Odia and willing to contribute parallel pair is an eligible member
  - All those members can tag parallel pair to the tweetbot by doing @[tweetbot_id]{Pair}
- Tweets representing present state of events, hence candidate pairs are relevant
- Variety of interests amongst the community: hence it has a healthy diversity in the candidate pairs
- Diversity in corpus ensures that the training data is close to real life
  - Specialized corpus in different domains can be extracted out if needed
- A few examples representing diversity
  - @Odiaculture tweets about Odia culture
  - @drgynec tweets poetry

Eligible Pairs (*Ep)*

- All pairs contributed by the community are candidate pairs
- From those candidate pairs, pairs are selected by moderators based on eligibility
- Eligibility is solely based on correctness of a pair
    - (A natively speaking individual with a graduate education makes a judgment of correctness by following the rules of grammar in both languages)
- All pairs generated by the moderator team are eligible
    - Eligibility follows the same rules as the eligibility of moderator pairs
    - The pairs generated by moderators are peer reviewed as an additional quality gate
- Moderation team follows do and do not list to pick candidate tweets eligible to generate pairs, the do and do not list is primarily used to maintain the positivity around the timeline.
    - Do and Do not list is a simple thumb rule to avoid moderators' human biases entering into the corpus un-internationally, and keep the stream only having parallel pair
    - One moderator cannot retweet their tagged pair into the stream
    - Simplicity and Objectivity is the driving factor for the do and don't lists
    - The do and do not list introduces a positive bias in the stream which is unnatural for a language corpus (negative language is also part of the language), however it is well considered and consideration possible to build a negative stream quickly as a separate project

Algorithms

- Regular expressions were used to match the style guide
    - Regular expressions are used based on the patterns mentioned above in the tweet style section
        - ■ <Beginning><English Sentence><separator><Translated Odia Sentence><End>
        - ■ <Beginning><Odia Sentence><separator><Translated English Sentence><End>
        - ■ <Beginning> <Translation Tweet> <Sub Tweet><End>

Toolset

- A twitter account

A twitter account was created as an interface with the Odia Twitterati. The account interacts with the community as a bot (using twitter API).

- A team of volunteer moderators

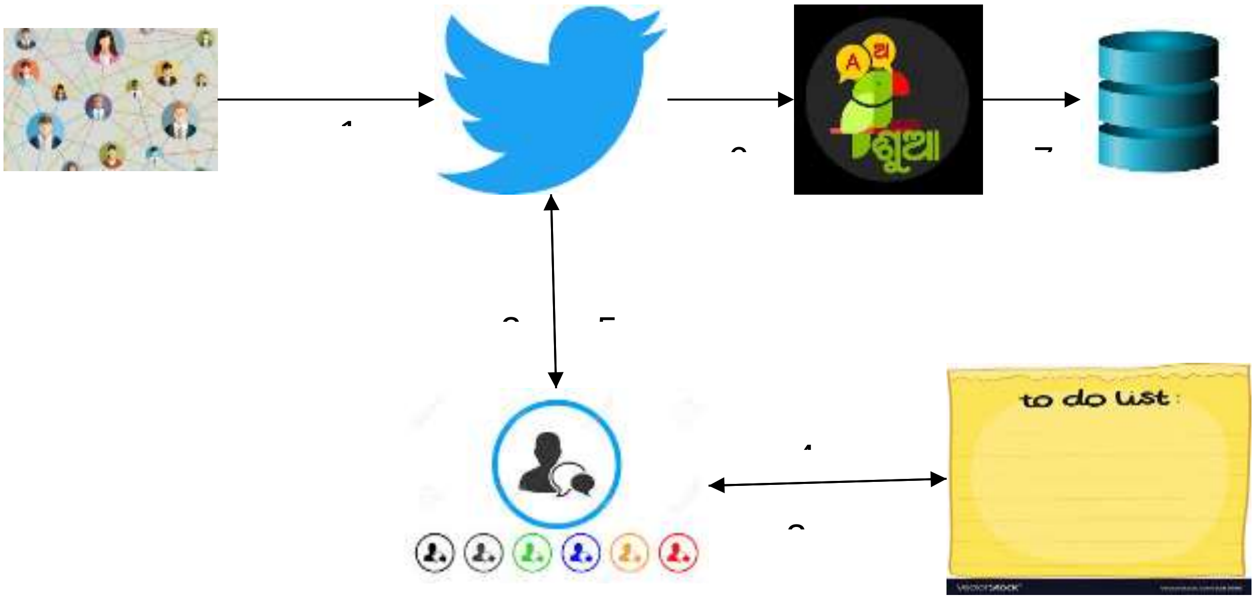
A team of volunteer moderators were created. An engagement policy was prepared, do and do not list was created as a guideline to the moderators. Simplicity, Objectivity, Goal Orientation (Generating parallel pairs), and Scalability (scaling the team size to create bigger streams) were the basis of defining the do and do not guides and engagement policy.

- Technology volunteer teams

A tech team of volunteers *created tools to read the twitter stream, parse the pairs, and store them in a storage* pluggable to a data pipeline which will feed the neural models that need them.

Another tech team of volunteers worked to *read that continuous stream of data and train the neural model.*

Process



- Open community contributions
    1. The community contributes pairs through twitter

- Moderation through a simple self-created to do and do not list
    - 2: Moderator reviews the pairs tagged
    - 3: By referring the do and do not lists
    - 4: makes a decision
    - 5: moderator re-tweets the tagged tweet
    - 6: the retweeted pair becomes part of the automatic stream

Applications

- This press model is usable by any system which needs a parallel corpus stream.
- This model is active and running in the form of a tweet bot named ଶୁଆ for Odia English parallel corpus.
- Similarly, this model can be applied to create any parallel pair stream for any language corpus computationally underrepresented Languages May involve the community in this way and build their own parallel corpus.
- Other language forms ranging from revered Sanskrit to marginalized Olchiki which are yet to be represented in modern NLP space like machine translation can also use this model to generate corpus streams.
- We urge Linguistics researchers and computational experts in the field of Natural Language Processing related computing to look at using this model for computationally underrepresented languages.
- There can be numerous NLP applications on it and it may even bring dead languages like Sanskrit alive, it may bring other underrepresented languages to the fore
- We recommend language activists to observe this model and derive approaches where this model can help their languages
- We also request a global collaboration for natural language processing in this age of advanced capabilities in computing and deep learning.

Extensions

- Twitter and tweet bot were chosen as the catchment platform for generating feed, however we believe that this model can be used on other social media platforms.

- For Facebook, it might be a stricter regular expression based parsing when applied it will enable long-range translations like paragraphs and pages.
- We urge that this model to be extended to platforms like Quora, Reddit etc., to be able to as big a stream possible for a parallel corpus needed for a specific language
- We also think that this model extension may be used and created at the undergraduate levels in universities in Odisha by the students coordinating with their professors and their Odia and English counterpart making it a multidisciplinary activity

Conclusion

The present analysis of models depicts how to generate a parallel corpus stream and description of methodology process and support of infrastructure to achieve the model. The choices made to apply these above models and formula to make an operational and functional viability. There are avenues of a few computer based applications that will open up a possibility to extent and use of social media platforms as catchment net. Therefore, Machine Translation from English to Odia is a need of the hour. The larger implications of these applications would be to find Odia language in any digital search engine translation board.

==

Glossary

- Active Learning: In Machine Learning, active learning is a process in which a machine, an agent, a computer program, a bot learns actively from the new data, this is in contrast to the models where Training is performed on available data only, next batch of data available need to wait for the next training cycle.
- Catchment platform: the platform which will act as a base from which parallel corpus will be mined, it has been derived from the "catchment area" in geology.
- Social Media: A suite of applications running on the internet Facebook, Twitter, LinkedIn, Snapchat, TikTok, WhatsApp which enables information to be passed from individual to individual with the society they know taking an active part in it.

- SOTA: State of the art: agreed by the research community to be the best available at that moment

References

Name	Link
The Effects of Corpus Size and Homogeneity on Language Model Quality	https://www.aclweb.org/anthology/W97-0118.pdf
One Billion Words for language modelling	https://opensource.google/projects/lm-benchmark
A Simple Introduction to Natural Language Processing	https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32
Natural Language Processing	https://en.wikipedia.org/wiki/Natural_language_processing
Parallel Corpora	http://www.ilc.cnr.it/EAGLES96/corpustyp/node20.html
Machine Translation	https://en.wikipedia.org/wiki/Machine_translation
Neural Machine Translation	https://en.wikipedia.org/wiki/Neural_machine_translation
Google Translate	https://translate.google.com/
Attention is All You Need	https://arxiv.org/abs/1706.03762
Compressive Transformers for long-range sequence modelling	https://openreview.net/attachment?id=SylKikSYDH&name=original_pdf
Twitter API	https://developer.twitter.com/en/docs/api-reference-index
Active Learning Theory and Applications	http://www.robotics.stanford.edu/~stong/papers/tong_thesis.pdf
Learning to Continually Learn	https://arxiv.org/abs/2002.09571
SOTA machine Translation papers	https://paperswithcode.com/task/machine-translation

==

Appendix

Sample parallel pairs from MTE2O ଶୁଆ timeline in JSON format

(JSON." https://www.json.org/. Accessed 8 Aug. 2020)

1. {od: “ନୌକାଟିଏ ପାଣି ଉପରେ ରହିବ, କିନ୍ତୁ ପାଣି ନୌକା ଭିତରେ ରହିବା ଉଚିତ ନୁହେଁ ; ତାହା ନୌକାଟିକୁ ନିଷ୍କଳ ଏବଂ ନିଷ୍ପ୍ରୟୋଜନ କରିଦେବ” ,en: “A boat will be on water , but water should not be in the boat ; that will make the boat stagnant and unfit for the purpose for which it is meant.” }
2. {en:”Old ways won't open new doors.”,od:”ପୁରୁଣା ପଦ୍ଧତିରେ ନୂଆ ଦୁଆର ଖୋଲିବ ନାହିଁ |”}
3. {od:”ହର-ପାର୍ବତୀଙ୍କ ହାତରେ କିଏ?”,en:”Who is in the hands of Har-Parvati?”}
4. {od:”ବରଗଡ଼ର ରାତି ହାଟ” ,en:”The Night Wholesaling Vegetable Market of Bargarh}
5. {en:”Proud moment for Odisha”,od:” ଓଡ଼ିଶା ପାଇଁ ଗର୍ବର ମୁହୂର୍ତ୍ତ"}
6. {'ନା' କିପରି କହିବେ କୌଣସି ଦୋଷ ଅନୁଭବ ନକରି? ,en:”How to SAY 'NO' without feeling any guilt?"}
7. {ଖରାପ ସଂଗତି, ସେଇ କୋଇଲା ଭଳି, ଯଦି ଗରମ ଥିବ ତ ହାତକୁ ଜଳେଇ ଦବ,ଆଉ ଯଦି ଥଣ୍ଡା ଥିବ ସେ ବି ହାତକୁ କଳା କରିଦେବ ||,en:”Bad company is like charcoal, if hot it will scald the hand that holds it, if cold it will make it black."}
8. {en:”A glass shaping process.”,od:”ଏକ କାଚ ଆକୃତି ତିଆରି ପ୍ରକ୍ରିୟା।"}
9. {en:”Quality is not an act, it is a habit.”,od:” ଗୁଣମାନ ଏକ କର୍ମ ନୁହେଁ, ଏହା ଏକ ଅଭ୍ୟାସ।"}
10. {en:”Patience is bitter, but its fruit is sweet. ”,od:”ଧୈର୍ଯ୍ୟ ହେଉଛି ପିତା, କିନ୍ତୁ ତା ଫଳଟି ମିଠା।"}

==

Anjan Kumar Panda, MSC IT, KSOU Mysore
Language Technology Activist, NLP Practitioner,
Internet Application Specialist, Technology Manager
Life Member, OSA. The Odisha Society of the Americas
5050, Hacienda Drive, Apt 2232, Dublin, CA, 94568
panda.anjankumar@gmail.com
Contact: 1- 845-535-0961

Dr Arun Kumar Malik, PhD, Assistant Professor of Political Science
Gujarat National Law University, Gandhinagar
amalik@gnlu.ac.in

Contact No. 8128650850