Computational Linguistics II: Comp Semantics and Discourse parsing
Project Report
Adequacy-Fluency Metrics for Machine Translation Evaluation

Karthikeyan Arumugam(2018900074)
Nikhil Bishnoi (2019114021)

Mentors : Vandan Mujadia and Hema Ala

Monsoon 2020

Contents

1 Introduction

Machine translation tasks always has challenges with automatic evaluation and their evaluation metrics. Many traditional MT-Evaluation metrics are unreliable and not capable enough to capture semantic level information. In particular Indian languages has very rich language structure and vocabularly. Some of our indian langugaes are free order and agglutinative in nature. This makes Evaluation process very difficult for any automatic evaluation schemes. On other hand human evaluators can provide and validate translation in semantic level. Since Human evaluation is time consuming and very expensive in nature , we need to create a evaluation metrics which can provide evaluation metrics computed based on semantic level information.

In our project we are exploring Adequacy-Fluency Metrics (AM-FM) for Machine translation. This AM-FM was proposed by Banchs R. E., D'Haro L.F., Li H. in "Adequacy - Fluency Metrics: Evaluating MT in the Continuous Space Model Framework", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol.23, No.3, pp.472482

Our Entire project work is to explore traditional evaluation metrics and compare it with AM-FM Evaluation metrics. We would like to produce experiment results which supports the claim that AMFm Evaluation metrics can provide better way of giving M.T evaluation

2 Related Works

- For Adequacy - Fluency Metrics , we found 2 published works.
- First journal is "Adequacy–fluency metrics: Evaluating MT in the continuous space model framework" by Banchs, Rafael E and D'Haro, Luis F and Li, Haizhou. published in 2015 - IEEE/ACM Transactions on Audio,Speech and language processing
- Second journal is "Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics" by D'Haro, Luis Fernando and Banchs, Rafael E and Hori, Chiori and Li, Haizhou. published in 2019 - Elsevier Computer Speech & Language

3 Data Collection and Processing

- For this project we have experimented with English-Hindi and English-Tamil Langauge pair
- Implementation and testing was done with Data used in WAT2019 competition.
  (http://lotus.kuee.kyoto-u.ac.jp/WAT/)
- For English-Hindi
  - Data source is IIT Bombay (IITB)
  - Link : http://www.cfilt.iitb.ac.in/iitb_parallel/
  - Size of corpus - 1589K Sentence pair .
  - Size of Dev Sub set which is used for translation experiments - 520 Sentence Pair
  - svm pretrained model weights from original paper - Link mentioned in reference section
  - language model built using original 1589K sentence pair - pretrained values reused from original paper - Link mentioned in refernce section
  - testing done with dev version of the en-hi pair 520 Sentences
- For English-Tamil
  - Data source is Charles University - English Tamil parallel corpus
  - Link : http://ufal.mff.cuni.cz/ ramasamy/parallel/html/
  - Size of corpus - 166K Sentence pair .
  - Size of Dev Sub set which is used for translation experiments - 500 Sentence Pair
  - svm pretrained model weights from original paper - Link mentioned in reference section - trained on 166K sentence Pair
  - language model built using original 166K sentence pair - pretrained on values reused from original paper - Link mentioned in refernce section
  - testing done with dev version of the en-ta pair 500 Sentences
- Data set links are mentioned in the reference section , a copy of the data is available in google drive folder shared in the reference section

4 Experiments

For this project , we have devised three levels of experiment. Each level experiments are done in English-Hindi and English-Tamil Language Pair. All experiments are implemented in python. Some of the model weights are pretrained . On of the original Author of AM-FM method - Luis Fernando D'Haro (lfdharo@die.upm.es) shared it with us over email.

- First Phase of Experiment :
  In the first phase od experiment , we generated MT translated output for the Data set we collected.Due to large volume of data, We couldn't manually process MT results from sampark/anusaaraka or other systems. Hence we used Microsoft's Cognitive Service from Azure Cloud to Translate the data in batches.

  Code for Microsoft's cognitice service based MT is available in our github repository.

- Second Phase of Experiment :
  Building MT Evaluation systems, To compare and analyse results , we coded traditional evaluation metrics using python , some of them are based upon nltk/sklearn libraries. We used pretrained weights for SVM and language model for en-hi/en-ta language pairs. We build a Standalone AM-FM Evaluation in python and used results of First phase to compare traditional and AM-FM metrics. All Translations are done in sentence level. Evaluation metrics are measured for each sentence pairs.

- Third Phase of Experiement :
  We collated results of both AM-FM evaluation and Traditional evaluation metrics. We compared their results and reviewed all metrics scores and provided qualitative analysis for Key examples.

5 Major Challenges

- Requirement of Huge Corpus to build AM-FM models which evaluates MT-metrics

- For Asian and indian languages - Models were build on unicode character level. Hence if the test evaluation is done with different Unicode characters for the language then AM-FM might not work as expected
  Example : for Tamil there are 29 unicoder (Refer *https://github.com/Ezhil-Language-Foundation/open-tamil/blob/master/tamil/txt2unicode/README.md*). the most used on in web is tscii unicode. Hence this could be potential problem

- Current implementation of AM-FM is completed for monolingual Adeuacy Calculation. Implementation of Cross language Adequacy requires much more complicated system and more parallel data.

- AM-FM Evaluation model build is time consuming. traditional metrics are light weight components when compared to AM-FM metrics.

6 Observations

- AM-FM Metrics were able to capture scenarios such as different word order, Sense relations such as Synonymy, polysemy, hyponymy etc better than traditional metrics
- AM-FM - system can provide proper computation only when we have large quantity of corpus to train the language models and Term Document matrix which can provide joint distribution of words , documents.
- for Domain Specific MT tasks, we need to build corpus on the given domain before we build AM-FM model weights. Domain specific Sense information can be learned only when we have domain specific Corpus.
- In our experiments , even when there are scenarios where results had different sentence structure or word order, AM - FM metrics were able to grasp those infromation much better than other evaluation metrics

7 Scope for further work

- In Future we would like to explore - Domain specific AM-Fm Evaluation metrics.
- Currently there are new Deep learning based AM-FM Metric systems are being built.
  We can Extend current system into a deep learning based model
  Refer : *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5)*

8 Conclusion

Proposed Adequacy-Fluency Metrics (AM-FM) framework provides better and more relevant evaluation score to the MT Task.
Mono lingual - Adequacy-Fluency Metrics (mAM-FM) provides superior results when compared to traditional evaluation metrics.
mAM-FM can provide better correlated assessments when compared to human evaluations.

9 Resources

- IITB - Corpus Link
  http://www.cfilt.iitb.ac.in/iitb_parallel/

- Charles University EN-TA Corpus Link
  http://ufal.mff.cuni.cz/ ramasamy/parallel/html/

- AI4Bharath resourse page
  https://github.com/AI4Bharat

- Github Repository Link
  https://github.com/karumugamio/CL2_AM_FM_Metrics_Analysis

- Our Project Google Drive Folder link
  https://drive.google.com/drive/folders/1D3n-U2REUdTP7nSUkJ9tvbQF6EJdnwm0?

- Resources shared by one of Author of AM-FM Framework

  – Model files for EN-HI (Build on IITB -EN-HI Corpus) -
    https://drive.google.com/drive/folders/1BawAIb09T-Mg4ez-HpLBpefWCdHKuUPX?

  – Model files for EN-TA - (Build on Charles University Corpus as WAT2019 Workshop) -
    https://drive.google.com/drive/folders/1zjuAgcWc6tZdimNAi197shBbaHl3T0ts?

  – Deep - AM-FM Model Data
    https://drive.google.com/drive/folders/1roT2HwmVIZtY1NdxTvfuj5iO1uctOeda?

  – new Deep learning based AM-FM Metric systems which are being built.
    Refer : *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation (https://link.springer.com/chapter/10.1007/978-981-15-8395-7_5)*

10 References

1. Adequacy–fluency metrics: Evaluating MT in the continuous space model framework,
   Banchs, Rafael E and D'Haro, Luis F and Li, Haizhou,
   IEEE/ACM Transactions on Audio, Speech, and Language Processing
   v23,IEEE 2015

2. Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics
   D'Haro, Luis Fernando and Banchs, Rafael E and Hori, Chiori and Li, Haizhou
   Computer Speech & Language Journal 2019 Elsevier v55,2019

3. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages

Anoop Kunchukuttan and Divyanshu Kakwani and Satish Golla and Gokul N.C. and Avik Bhattacharyya and Mitesh M. Khapra and Pratyush Kumar
2020-arXiv preprint arXiv:2005.00085

4. Presentation by Haizhou Li - ”https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2015/talk/IT_2.pdf”

5. Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007), (Meta-) evaluation of machine translation, in Proceedings of Statistical Machine Translation Workshop, pp. 136-158

6. C. Tillmann et al., “Accelerated DP Based Search for Statistical Translation”, in Proc. of the 5th European Conf. on Speech Commun. and Tech., Rhodos, Greece, Sept 1997, pp. 2667–2670.

7. K. Papineni et al., “BLEU: a method for automatic evaluation of machine translation”, in Proc. of the 40th Annu. Meeting of the Assoc. for Computational Linguistics, Philadelphia, PA, USA, Jul 2002, pp. 311-318

8. G. Doddington, “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics”, in Proc. of the Human Lang. Tech. Conf., San Diego, CA, USA, Mar 2002

9. A. LavieandM.J. Denkowski, “The Meteor metric for automatic evaluation of machine translation”, Machine Translation, vol. 23, pp. 105-115, May 2009

10. M. Snoveret al., “Study of Translation Edit Rate with Targeted Human Annotation”, in Proc. of the 7th Biennial Conf. of the Assoc. for Mach. Translation in the Amer., Cambridge, MA, USA, Aug 2006

11. C.K. Lo and D. Wu, “MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on seman-tic roles”,in Proc. ofthe49th Annu. Meeting oftheAssoc. forComputationalLinguistics, Portland, OR, USA, Jun 2011, pp. 220-229

11 Appendix a - Qualitative Analysis - result examples

1. Example 1 : Different Word order and minorly varient sandhi results.
   Given English Sentence : The government and the military blame the LTTE for keeping civilians against their will as ”human shields.”

   Reference Tamil Sentence :
   புலிகள் பொது மக்களை அவர்களது விருப்பதற்கு எதிராக "மனிதக் கேடயங்களாக" வைத்திருப்பதாக அரசாங்கமும் இராணுவமும் குற்றஞ்சாட்டுகின்றன.
   Romaized : puligal - pothumakkal avargalathu viruparthirku yethira ga manidha kedayamngalaga vaithu irupathaga arasangamum ranuvamum kutram saatu gindrana

   Microsoft translated Sentence :
   பொதுமக்களை அவர்களின் விருப்பத்துக்கு எதிராக "மனிதக் கேடயங்கள்" என்று புலிகள் வைத்திருப்பதாக அரசாங்கமும் இராணுவமும் குற்றம் சாட்டுகின்றன. .

   Romanized : pothumakkalai avargalil viruparthuku yethiraga manitha kedayangalaga yendru puligal vaithu irupathaga arasangamum raanuvamum kuttram saathu gindrana.

   AM -FM Score : mAmFm - 0.95 (calculated with alpha 0.5) ; am - 0.96 ; fm - 0.93

   Traditional Score : Meteor - 0.46 ; Bleu- 0.8 ;

2. Example 2 : Active vs Passive Voice comparison
   Given English : Our best wishes to Ranjitha! Reference Tamil Sentence :
   வாழ்த்துக்கள் ரஞ்சிதா!

   Romanized : Valthukal Ranjitha
   Microsoft Translated Sentence : ரஞ்சிதாவுக்கு எங்கள் வாழ்த்துக்கள்!

   Romanized : Ranjithavukku yengal valthukal

   AM -FM Score : mAmFm - 0.672 (calculated with alpha 0.5) ; am - 0.89 ; fm - 0.45

   Traditional Score : Meteor - 0 ; Bleu- 0.5 ;

3. Example 3 : Code mix/Tanglish words in reference sentence
   Given English sentence: That’s a good policy. Keep it up!
   Reference tamil sentence : "நல்ல பாலிஸி, தொடரட்டும்!

Romanized : Nall policy thodaratum.
Microsoft translated tamil sentence : அது ஒரு நல்ல கொள்கை. அதை வைத்து!
Romanized : athu oru nalla kolgai , athai vaithu.

AM -FM Score : mAmFm - 0.2 (calculated with alpha 0.5) ; am - 0.39 ; fm - 0.009

Traditional Score : Meteor - 0.15 ; Bleu- 0.14 ;

4. Example 4 : Phoentically different spelling translations

   Given English Sentence : His ’Nishabd’ has been inspired by ’Lolita.’

   Reference tamil sentence : 'லோலிட்டா' படத்தை தழுவி 'நிஷப்த்' படத்தை இயக்கினார்.
   Romanized : lolita padathai thaluvi nisabth padathai iyakki naar

   Microsoft Translated Sentence: அவரது 'நிஷாப்த்' 'லோலிதா' ஈர்க்கப்பட்டு உள்ளது.
   romanized : avarathu nisabth lolitha eerkapattu ullathu

   AM -FM Score : mAmFm - 0.46 (calculated with alpha 0.5) ; am - 0.63 ; fm - 0.29

   Traditional Score : Meteor - 0 ; Bleu- 0.32 ;