R & D Project I

Presented by

Suraj Kumar (18305R008)

Guided by

Prof. Pushpak Bhattacharyya

June 24, 2020

Outline

- Textual Entailment
- Paraphrasing
- Machine Translation Evaluation
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- Results and Observations
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

Outline

- **Textual Entailment**
- Paraphrasing
- Machine Translation Evaluation
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- Results and Observations
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

Textual Entailment

Relationship between premise (p) and hypothesis (h) indicating the inclusion of meaning of h in p

p: *"If you join the Army, you will serve the country."*
h: *"Join the Army to serve the country."*

p: *"If you join the Army, you will serve the country."*
h: *"You cannot serve the country in any way."*

p: *"If you join the Army, you will serve the country."*
h: *"Be there for the country in difficult times."*

Language Features for TE

1. **Lexical**
    a. N-grams matching
    b. Levenshtein distance

    p : *"Since it was raining, I didn't go to school"*
    h : *"I didn't go to school"*

2. **Syntactic**
    a. Dependency Tree

    p : *"Microsoft bought Linkedin"*
    h : *"Microsoft owns Linkedin"*

3. **Semantic**

    p: *"In 1948 Nathuram Godse murdered Mahatma Gandhi"*
    h : *"Mahatma Gandhi died in 1948"*

Challenging!

Outline

- Textual Entailment
- **Paraphrasing**
- Machine Translation Evaluation
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- Results and Observations
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

Paraphrasing

Analyzing sentences that are semantically identical

p: *"There were many people in the ML class yesterday"*

q: *"Huge number of people turned up for the ML class yesterday"*

Paraphrases can be detected using bi-directional entailment relationship

Applications:

1. Question-Answering
2. Text Summarization
3. Machine Translation Evaluation
4. Data Augmentation

Outline

- Textual Entailment
- **Paraphrasing**
- **Machine Translation Evaluation**
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- Results and Observations
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

Machine Translation Evaluation (MTE)

- Measures the quality (Adequacy and/or Fluency) of translation systems by comparing reference and candidate translation
    - Both reference and candidate sentences should be semantically equivalent for good quality system
- Two methods for MTE:
    - Human Evaluation
    - Automatic Evaluation

Why Automatic MTE ?

1. Inexpensive
2. Less time to evaluate MT outputs than humans
3. Monitor incremental system changes during development
4. Fast comparison of different systems performance

E.g. Precision, Recall, BLEU, METEOR, LAYERED etc

Outline

- Textual Entailment
- Paraphrasing
- Machine Translation Evaluation
- **Language Representation / Embedding**
- MT evaluation using Bi-directional Entailment
- Results and Observations
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

Language Representation

- Language representation is an important aspect towards applying deep learning models to NLP
- Experimented with 5 language representations
  1. BERT
  2. XLNet
  3. RoBERTa
  4. ALBERT
  5. XLM

BERT

BERT is based on Transformer Network (Encoder)

Two tasks are used for pre-training of BERT:

1. Masked Language Modelling (MLM)
2. Next Sentence Prediction (NSP)

Masked LM

Sentence: I love to code in python

Masked Sentence: I [MASK] to code in python

[MASK] means token is missing

Task: Predict the [MASK] (in this case [MASK] will be love)

Goal: To understand relationship between words

Note: We randomly choose 15% words in each sentence and replace it with:

1. [MASK] token 80% of the time
2. A random token 10% of the time
3. Unchanged 10% of the time

Next Sentence prediction

Goal: To understand the relationship between sentences

Given 2 sentences A & B:

Task: Is B the actual next sentence in the corpus (Simple binary classification

with 2 labels **IsNext** and **IsNotNext**)

Note: Specifically, when choosing the sentences A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A (labeled as IsNext), and 50% of the time it is a random sentence from the corpus (labeled as IsNotNext)

RoBERTa

It does following modifications to BERT:

1. Training the model **longer**, with **bigger batches**, over **more data**
2. Removing the next sentence prediction objective
3. Training on **longer sequences**
4. Dynamically changing the masking pattern applied to the training data

ALBERT

Focuses on a very basic question: Is having better NLP models as easy as having larger models?

1. **Memory limitations** of available hardware
2. Training speed can also be significantly hampered in **distributed training,** as the communication overhead is directly proportional to the number of parameters in the model

XLM

Three objectives of XLM are:

1. Causal Language Modeling (CLM)

    Next Token Prediction

2. Masked Language Modeling (MLM)

3. Translation Language Modeling (TLM)

    Extension of MLM where instead of considering monolingual text streams, we concatenate parallel sentences. It randomly mask words in both the source and target sentences.

XLNet

Autoencoding Language Modelling (BERT, RoBERTa etc) has following drawbacks:

1. Does not perform **explicit density estimation**
2. **Pretrain-Finetune** discrepancy
3. Assumes the predicted masked tokens are **independent** of each other given the unmasked tokens

XLNet addresses all above issues

Maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order (permutation language modelling). In expectation, each position learns to utilize contextual information from all positions, i.e., capturing bidirectional context.

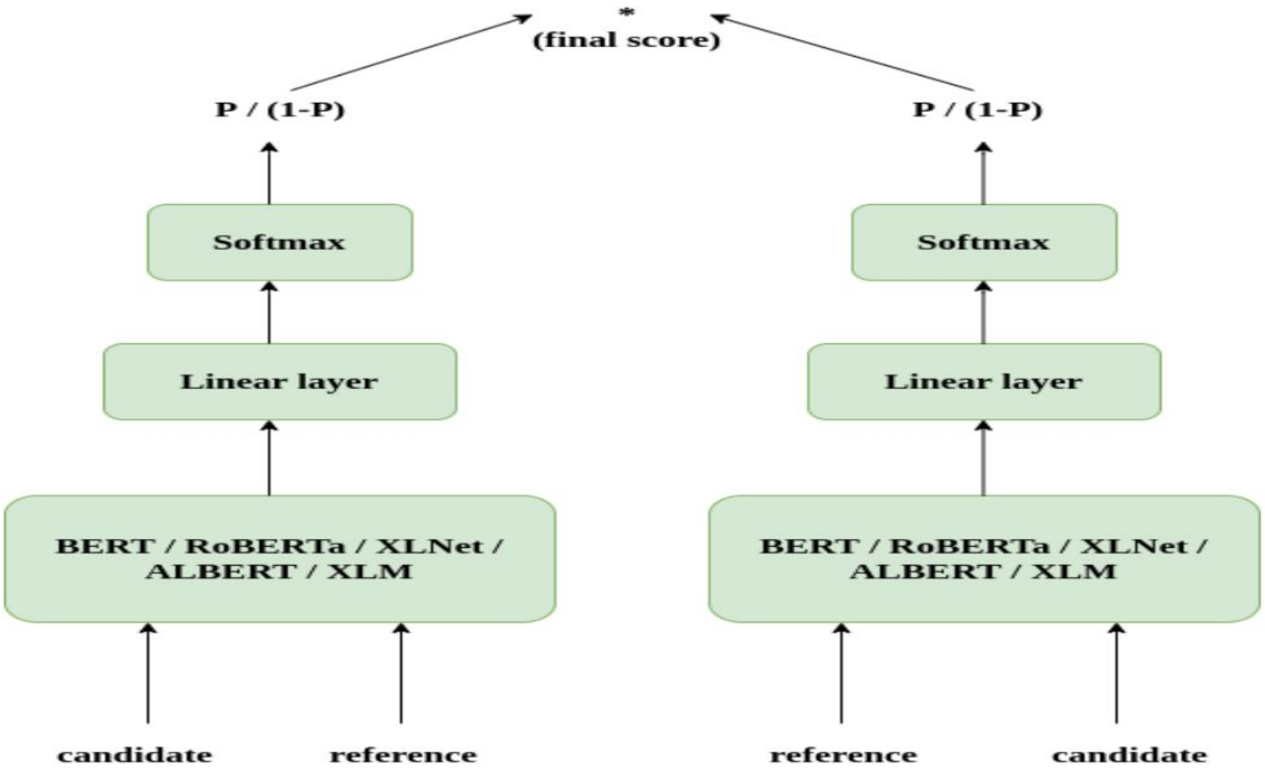
Outline

- Textual Entailment
- Paraphrasing
- Machine Translation Evaluation
- Language Representation / Embedding
- **MT evaluation using Bi-directional Entailment**
- Results and Observations
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

MT evaluation using Bi-directional Entailment

- Find if the reference and candidate sentences are same in meaning i.e. if they are **paraphrases** of each other
  - By checking if the entailment relationship holds between the candidate and a reference sentences in the forward and backward direction
- Used **Deep Learning Models** with 5 different pre-trained language representations
- Fine-tuned on **MNLI** dataset (433k sentence pairs)

Architecture



Outline

- Textual Entailment
- Paraphrasing
- Machine Translation Evaluation
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- **Results and Observations**
- Conclusion and Future Work
- Demo & EMNLP 2020 Paper

Results & Observations

- Evaluated on both **system** and **segment** level using following data:
  1. WMT 14
  2. WMT 17
  3. WMT 18
  4. WMT 19 (only system level)

These results are part of **EMNLP 2020** paper titled "Machine Translation Evaluation Using Bi-directional Entailment"

System Level Evaluations

WMT 14

Metric	fr-en	de-en	hi-en	cs-en	ru-en	Average	SpearAvg
DiscoTK-party-tuned	0.977	0.943	0.956	0.975	**0.870**	**0.944**	**0.912**
LAYERED	0.973	0.893	**0.976**	0.941	0.854	0.927	0.894
DiscoTK-party	0.970	0.921	0.862	0.983	0.856	0.918	0.856
UPC-STOUT	0.968	0.915	0.898	0.948	0.837	0.913	0.901
VERTa-W	0.959	0.867	0.920	0.934	0.848	0.906	0.868
VERTa-EQ	0.959	0.854	0.927	0.938	0.842	0.904	0.857
tBLEU	0.952	0.832	0.954	0.957	0.803	0.900	0.841
BLEU_NRC	0.953	0.823	0.959	0.946	0.787	0.894	0.855
BLEU	0.952	0.832	0.956	0.909	0.789	0.888	0.833
UPC-IPA	0.966	0.895	0.914	0.824	0.812	0.882	0.858
APAC	0.963	0.817	0.790	0.982	0.816	0.874	0.807
REDSys	**0.981**	0.898	0.676	0.989	0.814	0.872	0.786
REDSysSent	0.980	0.910	0.644	**0.993**	0.807	0.867	0.771
NIST	0.955	0.811	0.784	0.983	0.800	0.867	0.824
CDER	0.954	0.823	0.826	0.915	0.802	0.864	0.807
DiscoTK-light	0.965	0.935	0.557	0.954	0.791	0.840	0.774
Meteor	0.975	0.927	0.457	0.980	0.805	0.829	0.788
TER	0.951	0.772	0.616	0.989	0.810	0.827	0.746
WER	0.952	0.762	0.619	0.992	0.809	0.827	0.736
PER	0.946	0.867	0.432	0.937	0.799	0.796	0.758
AMBER	0.948	0.910	0.506	0.744	0.797	0.781	0.728
ELEXR	0.971	0.857	0.535	0.945	-0.404	0.581	0.652
BiDiEnt.BERT	0.979	**0.963**	0.926	0.904	0.843	0.923	**0.912**
BiDiEnt.XLNet	0.946	0.946	0.889	0.654	0.724	0.832	0.831
BiDiEnt.ALBERT	0.827	0.716	0.714	0.939	0.796	0.798	0.799
BiDiEnt.XLM	0.903	0.763	0.421	0.688	0.825	0.720	0.498
BiDiEnt.RoBERTa	-0.350	0.129	-0.426	-0.120	0.144	-0.124	-0.058

BERT language representation gives the best result among all language representations followed by **XLNet** which is also very competitive to BERT except for cs-en and ru-en pair.

WMT 17

Metric	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
AutoDA	0.438	0.959	0.925	0.973	0.907	0.916	0.734
BEER	0.972	0.960	0.955	0.978	0.936	0.972	0.902
Blend	0.968	**0.976**	0.958	0.979	**0.964**	0.984	0.894
BLEU	0.971	0.923	0.903	0.979	0.912	0.976	0.864
bleu2vec_sep	0.989	0.936	0.888	0.966	0.907	0.961	0.886
CDER	0.989	0.930	0.927	0.985	0.922	0.973	0.904
CharacTER	0.972	0.974	0.946	0.932	0.958	0.949	0.799
chrF	0.939	0.968	0.938	0.968	0.952	0.944	0.859
chrF++	0.940	0.965	0.927	0.973	0.945	0.960	0.880
MEANT_2.0	0.926	0.950	0.941	0.970	0.962	0.932	0.838
MEANT_2.0-nosrl	0.902	0.936	0.933	0.963	0.960	0.896	0.800
ngram2vec	0.984	0.935	0.890	0.963	0.907	0.955	0.880
NIST	**1.000**	0.931	0.931	0.960	0.912	0.971	0.849
PER	0.968	0.951	0.896	0.962	0.911	0.932	0.877
TER	0.989	0.906	0.952	0.971	0.912	0.954	0.847
TreeAggreg	0.983	0.920	**0.977**	0.986	0.918	**0.987**	0.861
UHH_TSKM	0.996	0.937	0.921	**0.990**	0.914	**0.987**	0.902
WER	0.987	0.896	0.948	0.969	0.907	0.925	0.839
BiDiEnt.BERT	0.997	0.916	0.930	0.952	0.946	0.909	0.853
BiDiEnt.XLNet	0.982	0.960	0.970	0.978	0.957	0.907	**0.929**
BiDiEnt.ALBERT	0.911	0.927	0.231	0.699	0.833	0.870	0.608
BiDiEnt.XLM	0.842	0.686	0.381	0.835	0.871	0.903	0.480
BiDiEnt.RoBERTa	0.129	0.113	0.786	0.805	0.281	0.860	0.302

XLNet language representation gives the best result among all language representations followed by **BERT** which is also very competitive to XLNet.

WMT 18

Metric	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
BEER	0.958	0.994	0.985	0.991	0.982	0.870	0.976
BLEND	0.973	0.991	0.985	0.994	0.993	0.801	0.976
BLEU	0.970	0.971	0.986	0.973	0.979	0.657	0.978
CDER	0.972	0.980	0.990	0.984	0.980	0.664	**0.982**
CharacTER	0.970	0.993	0.979	0.989	0.991	0.782	0.950
chrF	0.966	0.994	0.981	0.987	0.990	0.452	0.960
chrF+	0.966	0.993	0.981	0.989	0.990	0.174	0.964
ITER	0.975	0.990	0.975	**0.996**	0.937	0.861	0.980
meteor++	0.945	0.991	0.978	0.971	0.995	0.864	0.962
NIST	0.954	0.984	0.983	0.975	0.973	0.970	0.968
PER	0.970	0.985	0.983	0.993	0.967	0.159	0.931
RUSE	**0.981**	**0.997**	0.990	0.991	0.988	0.853	0.981
TER	0.950	0.970	0.990	0.968	0.970	0.533	0.975
UHH_TSKM	0.952	0.980	0.989	0.982	0.980	0.547	0.981
WER	0.951	0.961	**0.991**	0.961	0.968	0.041	0.975
YiSi-0	0.956	0.994	0.975	0.978	0.988	0.954	0.957
YiSi-1	0.950	0.992	0.979	0.973	0.991	0.958	0.951
YiSi-1_srl	0.965	0.995	0.981	0.977	0.992	0.869	0.962
BiDiEnt.BERT	0.973	0.991	0.971	0.946	0.970	**0.989**	0.961
BiDiEnt.XLNet	0.655	0.990	0.982	0.958	**0.999**	0.975	0.884
BiDiEnt.ALBERT	0.946	0.983	0.853	0.838	0.925	0.969	0.810
BiDiEnt.XLM	0.961	0.694	0.975	0.110	0.942	0.711	0.817
BiDiEnt.RoBERTa	0.487	0.933	0.900	0.601	0.719	0.810	0.118

BERT language representation gives the best result among all language representations followed by **XLNet** which is also very competitive to BERT except for cs-en pair

WMT 19

Metric	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
BEER	0.906	0.993	0.952	0.986	0.947	0.915	0.942
BERTr	0.926	0.984	0.938	0.990	0.948	0.971	0.974
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
CDER	0.890	0.988	0.876	0.967	0.975	0.892	0.917
CharacTER	0.898	0.990	0.922	0.953	0.955	0.923	0.943
chrF	0.917	0.992	0.955	0.978	0.940	0.945	0.956
chrF+	0.916	0.992	0.947	0.976	0.940	0.945	0.956
EED	0.903	0.994	0.976	0.980	0.929	0.950	0.949
ESIM	0.941	0.971	0.885	0.986	**0.989**	0.968	**0.988**
hLEPORa_baseline	—	—	—	0.975	—	—	0.947
hLEPORb_baseline	—	—	—	0.975	0.906	—	0.947
ibm1.morpheme	0.345	0.740	—	—	0.487	—	—
ibm1.pos4gram	0.339	—	—	—	—	—	—
LASIM	0.247	—	—	—	—	0.310	—
LP.1	0.474	—	—	—	—	0.488	—
Meteor.._2.0.syntax.	0.887	0.995	0.909	0.974	0.928	0.950	0.948
Meteor.._2.0.syntax.copy.	0.896	0.995	0.900	0.971	0.927	0.952	0.952
NIST	0.813	0.986	0.930	0.942	0.944	0.925	0.921
PER	0.883	0.991	0.910	0.737	0.947	0.922	0.952
PReP	0.575	0.614	0.773	0.776	0.494	0.782	0.592
sacreBLEU.BLEU	0.813	0.985	0.834	0.946	0.955	0.873	0.903
sacreBLEU.chrF	0.910	0.990	0.952	0.969	0.935	0.919	0.955
TER	0.874	0.984	0.890	0.799	0.960	0.917	0.840
UNI	0.846	0.930	—	—	—	0.805	—
UNI.	0.850	0.924	—	—	—	0.808	—
WER	0.863	0.983	0.861	0.793	0.961	0.911	0.820
WMDO	0.872	0.987	0.983	**0.998**	0.900	0.942	0.943
YiSi-0	0.902	**0.993**	**0.993**	0.991	0.927	0.958	0.937
YiSi-1	0.949	0.989	0.924	0.994	0.981	**0.979**	0.979
YiSi-1_srl	**0.950**	0.989	0.918	0.994	0.983	0.978	0.977
YiSi-2	0.796	0.642	0.566	0.324	0.442	0.339	0.940
YiSi-2_srl	0.804	—	—	—	—	—	0.947
BiDiEnt.BERT	0.920	0.969	0.900	0.929	0.972	0.972	0.977
BiDiEnt.XLNet	0.838	0.964	0.846	0.937	0.958	0.959	0.933
BiDiEnt.ALBERT	0.500	0.442	0.391	0.659	0.805	0.599	0.826
BiDiEnt.XLM	0.198	0.503	0.198	0.392	0.247	0.327	0.254
BiDiEnt.RoBERTa	0.767	0.855	0.270	0.659	0.758	0.692	0.721

BERT language representation gives the best result among all language representations followed by **XLNet** which is also very competitive to BERT.

Segment Level Evaluations

WMT 14

Metric	fr-en	de-en	hi-en	cs-en	ru-en	Average
DiscoTK-party-tuned	**0.433**	**0.380**	0.434	**0.328**	**0.355**	**0.386**
BEER	0.417	0.337	**0.438**	0.284	0.333	0.362
REDcombSent	0.406	0.338	0.417	0.284	0.336	0.356
REDcombSysSent	0.408	0.338	0.416	0.282	0.336	0.356
Meteor	0.406	0.334	0.420	0.282	0.329	0.354
REDSysSent	0.404	0.338	0.386	0.283	0.321	0.346
REDSent	0.403	0.336	0.383	0.283	0.323	0.345
UPC-IPA	0.412	0.340	0.368	0.274	0.316	0.342
UPC-STOUT	0.403	0.345	0.352	0.275	0.317	0.338
VERTa-W	0.399	0.321	0.386	0.263	0.315	0.337
VERTa-EQ	0.407	0.315	0.384	0.263	0.312	0.336
DiscoTK-party	0.395	0.334	0.362	0.264	0.305	0.332
AMBER	0.367	0.313	0.362	0.246	0.294	0.316
BLEU_NRC	0.382	0.272	0.322	0.226	0.269	0.294
sentBLEU	0.378	0.271	0.301	0.212	0.263	0.285
APAC	0.364	0.271	0.288	0.198	0.276	0.279
BiDiEnt.BERT	0.253	0.243	0.318	0.241	0.248	0.261
DiscoTK-light	0.311	0.224	0.238	0.187	0.209	0.234
DiscoTK-light-kool	0.005	0.001	0.000	0.002	0.001	0.002

fr-en: 0.18
de-en: 0.14
hi-en: 0.12
cs-en: 0.09
ru-en: 0.10

Min diff: 0.09

Max diff: 0.18

WMT 17

Metric	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
AutoDA	0.499	0.543	0.673	0.533	0.584	0.625	0.583
BEER	0.511	0.530	0.681	0.515	0.577	0.600	0.582
Blend	**0.594**	**0.571**	**0.733**	0.577	**0.622**	**0.671**	**0.661**
bleu2vec_sep	0.439	0.429	0.590	0.386	0.489	0.529	0.526
chrF	0.514	0.531	0.671	0.525	0.599	0.607	0.591
chrF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593
MEANT_2.0	0.578	0.565	0.687	**0.586**	0.607	0.596	0.639
MEANT_2.0-nosrl	0.566	0.564	0.682	0.573	0.591	0.582	0.630
ngram2vec	0.436	0.435	0.582	0.383	0.490	0.538	0.520
sentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512
TreeAggreg	0.486	0.526	0.638	0.446	0.555	0.571	0.535
UHH_TSKM	0.507	0.479	0.600	0.394	0.465	0.478	0.477
BiDiEnt.BERT	0.428	0.438	0.486	0.428	0.462	0.538	0.457

cs-en: 0.17
de-en: 0.13
fi-en: 0.24
lv-en: 0.16
ru-en: 0.16
tr-en: 0.13
zh-en: 0.20

Min diff: 0.13

Max diff: 0.24

WMT 18

Metric	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
BEER	0.295	0.481	0.341	0.232	0.288	0.229	0.214
BLEND	0.322	0.492	**0.354**	0.226	0.290	0.232	0.217
CharacTER	0.256	0.450	0.286	0.185	0.244	0.172	0.202
chrF	0.288	0.479	0.328	0.229	0.269	0.210	0.208
chrF+	0.288	0.479	0.332	0.234	0.279	0.218	0.207
ITER	0.198	0.396	0.235	0.128	0.139	-0.029	0.144
meteor++	0.270	0.457	0.329	0.207	0.253	0.204	0.179
RUSE	**0.347**	**0.498**	0.368	**0.273**	**0.311**	**0.259**	**0.218**
sentBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178
UHH_TSKM	0.274	0.436	0.300	0.168	0.235	0.154	0.151
YiSi-0	0.301	0.474	0.330	0.225	0.294	0.215	0.205
YiSi-1	0.319	0.488	0.351	0.231	0.300	0.234	0.211
YiSi-1_srl	0.317	0.483	0.345	0.237	0.306	0.233	0.209
BiDiEnt.BERT	0.204	0.362	0.273	0.189	0.195	0.215	0.112

cs-en: 0.14
de-en: 0.14
et-en: 0.08
fi-en: 0.08
ru-en: 0.12
tr-en: 0.04
zh-en: 0.11

Min diff: 0.04

Max diff: 0.14

Analysis of results

1. Pair encoding of reference and candidate translations which gives better results
2. Contextual embedding
3. Segment level results are not so good because BERT fine-tuning objective does not involve pairwise losses.

Limitations

1. Scores generated from our metrics go to extremes because of **uni-directional entailment**
2. As of now, our metric can judge translations where **target language is English**

Outline

- Textual Entailment
- Paraphrasing
- Machine Translation Evaluation
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- Results and Observations
- **Conclusion and Future Work**
- Demo & EMNLP 2020 Paper

Conclusion

1. Machine Translation Evaluation can be done by determining if candidate translation is paraphrase of the reference translation
2. Pre-trained language representation like BERT and XLNet is useful for finding bi-directional entailment / paraphrasing

Future Work

Short Term:

Ensemble some lexical features like word overlap, edit distance and other distance metrics into our proposed metric to produce better results.

Long Term:

Extend our metric to evaluate translations from English to other language pairs

Outline

- Textual Entailment
- Paraphrasing
- Machine Translation Evaluation
- Language Representation / Embedding
- MT evaluation using Bi-directional Entailment
- Results and Observations
- Conclusion and Future Work
- **Demo & EMNLP 2020 Paper**

Demo / Paper

- Submitted paper to **EMNLP 2020**

  Link to paper: https://drive.google.com/drive/folders/15xfxELR2-xeieg2mOxzhd5vaxM5t2p_f?usp=sharing

- Demo Link: http://www.cfilt.iitb.ac.in/BiDiEnt/

Additional Work

Results for mBLEU vs BLEU

Language Pair	BLEU	mBLEU
HI-BN	23.6	27.39
BN-HI	26.17	27.42
HI-MR	25.4	28.63
MR-HI	36.12	37.74

References

1. Jacob Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv 2018
2. Yinhan Liu et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv July 2019
3. G. Lample et al., Cross-lingual Language Model Pretraining, arXiv Jan 2019
4. Z. Lan et al., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv Sept 2019
5. Z. Yang et al., XLNet: Generalized Autoregressive Pre-training for Language Understanding. arXiv June 2019
6. Jesse Dodge et al., Fine-Tuning Pretrained language Models: Weight Initializations, Data Orders, and early Stopping, arXiv, 15 Feb 2020
7. Qingsong Ma et al., Results of the WMT18 Metrics Shared Task, ACL 2018
8. Qingsong Ma et al., Results of the WMT19 Metrics Shared Task, ACL 2019
9. Sebastian Pado et al., Robust Machine Translation Evaluation with Entailment Features, ACL 2009
10. Sebastian Pado et al., Textual Entailment Features for Machine Translation Evaluation, ACL 2008
11. Rakesh Khobragade et al., Machine Translation Evaluation Using Bi-directional Entailment, arXiv 2019
12. Matous Machacek et al., Machine Translation Evaluation Using Bi-directional Entailment, ACL 2014
13. Ondrej Bojar et al., Results of the WMT14 Metrics Shared Task, ACL 2017
14. Matt Post, A Call for Clarity in Reporting BLEU Scores, arXiv 2018
15. Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, ACL 2002
16. Chris Callison-Burch, Miles Osborne & Philipp Koehn, Re-evaluating the Role of BLEU in Machine Translation Research, ACL 2006

Thank You...

Backup Slides

Correlation Coefficient for mBLEU vs BLEU

Language Pair	BLEU	mBLEU
BN-HI	0.3216706335	0.3391451924

1. mBLEU has slightly better correlation coefficient than BLEU
2. Direct Assessment (DA) of 100 sentences is done manually by me on the basis of adequacy only