Machine Translation Evaluation

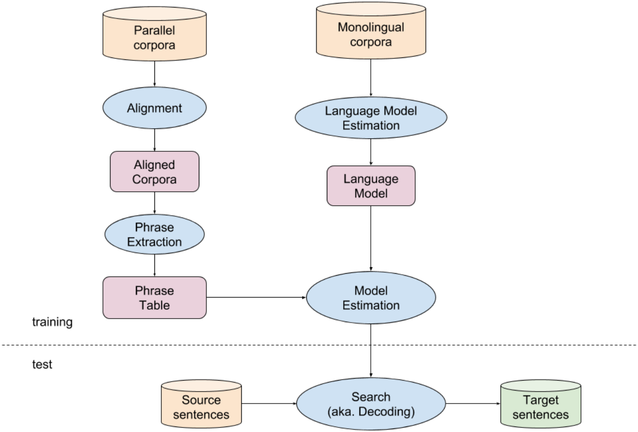
(Based on Miloš Stanojević's slides)

Iacer Calixto

Institute for Logic, Language and Computation
University of Amsterdam

May 18, 2018

Machine Translation Pipeline



"Good" versus "Bad" Translations

- How bad can translations be?

“Good” versus “Bad” Translations

- How bad can translations be?
  - Grammar errors:

“Good” versus “Bad” Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.

“Good” versus “Bad” Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.

"Good" versus "Bad" Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.

"Good" versus "Bad" Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.
  - Disfluent translations: e.g. She does not like [to] dance.

“Good” versus “Bad” Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.
  - Disfluent translations: e.g. She does not like [to] dance.
  - Etc.

"Good" versus "Bad" Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.
  - Disfluent translations: e.g. She does not like [to] dance.
  - Etc.
- What constitutes a good translation?

"Good" versus "Bad" Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.
  - Disfluent translations: e.g. She does not like [to] dance.
  - Etc.
- What constitutes a good translation?
  - One that accounts for all the "**units of meaning**" in the source sentence?

"Good" versus "Bad" Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.
  - Disfluent translations: e.g. She does not like [to] dance.
  - Etc.
- What constitutes a good translation?
  - One that accounts for all the "**units of meaning**" in the source sentence?
  - One that **reads fluently** in the target language?

"Good" versus "Bad" Translations

- How bad can translations be?
    - Grammar errors:
        - Wrong noun-verb agreement: e.g. She do not dance.
        - Spelling mistakes: e.g. The dog is playin with the bal.
        - Etc.
    - Disfluent translations: e.g. She does not like [to] dance.
    - Etc.
- What constitutes a good translation?
    - One that accounts for all the "**units of meaning**" in the source sentence?
    - One that **reads fluently** in the target language?
- What about translating literature, e.g. Alice's Adventures in Wonderland?

"Good" versus "Bad" Translations

- How bad can translations be?
  - Grammar errors:
    - Wrong noun-verb agreement: e.g. She do not dance.
    - Spelling mistakes: e.g. The dog is playin with the bal.
    - Etc.
  - Disfluent translations: e.g. She does not like [to] dance.
  - Etc.
- What constitutes a good translation?
  - One that accounts for all the "**units of meaning**" in the source sentence?
  - One that **reads fluently** in the target language?
- What about translating literature, e.g. Alice's Adventures in Wonderland?
- Or a philosophical treatise, e.g. Beyond Good and Evil?

Good Translations - Fluency vs. Adequacy

- Let's simplify the problem:
    - One axis of our evaluation should account for **target-language fluency**;

Good Translations - Fluency vs. Adequacy

- Let's simplify the problem:
  - One axis of our evaluation should account for **target-language fluency**;
  - Another axis should account for how **adequate** are the source-sentence "**units of meaning**" translated into the target language.

Good Translations - Fluency vs. Adequacy

- Let's simplify the problem:
  - One axis of our evaluation should account for **target-language fluency**;
  - Another axis should account for how **adequate** are the source-sentence "**units of meaning**" translated into the target language.
- Examples:
  - The man is playing football (source sentence)
  - La femme joue au football (✓ fluent but ✗ adequate)
  - ✗Le homme joue ✗football (✗ fluent but ✓ adequate)
  - L'homme joue au football (✓ fluent and ✓ adequate)

Why Machine Translation Evaluation?

- Why do we need automatic evaluation of MT output?

Why Machine Translation Evaluation?

- Why do we need automatic evaluation of MT output?
    - Rapid system development;
    - Tuning MT systems;
    - Comparing different systems;

Why Machine Translation Evaluation?

- Why do we need automatic evaluation of MT output?
  - Rapid system development;
  - Tuning MT systems;
  - Comparing different systems;
- Ideally we would like to incorporate **human feedback** too, but they are **too expensive**... ☹

What is a Metric?

- A **function** that computes the **similarity** between the output of an MT system (i.e. hypothesis or `sys`) and one or more human translations (reference translations or `ref`);

What is a Metric?

- A **function** that computes the **similarity** between the output of an MT system (i.e. hypothesis or `sys`) and one or more human translations (reference translations or `ref`);
- It can be interpreted in different ways:
  - Overlap between `sys` and `ref`: precision, recall...
  - Edit distance: insert, delete, shift;
  - Etc.

What is a Metric?

- A **function** that computes the **similarity** between the output of an MT system (i.e. hypothesis or `sys`) and one or more human translations (reference translations or `ref`);
- It can be interpreted in different ways:
  - Overlap between `sys` and `ref`: precision, recall...
  - Edit distance: insert, delete, shift;
  - Etc.
- Different metrics make different choices;

BLEU (Papineni et al., 2002)

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- Commonly, we set $N = 4$, $w_n = \frac{1}{N}$;
- BP stands for "Brevity Penalty" and is computed by:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- c is the length of the candidate translation;
- r is the effective reference corpus length.

BLEU (cont.)

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- `ref`: john plays in the park (length $= 5$)
- `hyp`: john is playing in the park (length $= 6$)
- `1-gram`: ✓john ✗is ✗playing ✓in ✓the ✓park
- $\text{BP} = 1$ $(c > r)$
- For $N = 1$:
  - $w_1 = \frac{1}{1} = 1$
  - $p_1 = \frac{4}{5}$, therefore $\text{BLEU}_1 = 1 \cdot \exp(1 \cdot \log 0.8) = 0.9$.

BLEU (cont.)

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- `ref`: john plays in the park (length = 5)
- `hyp`: john is playing in the park (length = 6)
- `1-gram`: ✓john ✗is ✗playing ✓in ✓the ✓park
- `2-gram`: ✗john is, ✗is playing, ✗playing in, ✓in the, ✓the park
- $\text{BP} = 1$ $(c > r)$
- For $N = 2$:
  - $w_1 = w_2 \frac{1}{2} = 0.5$
  - $p_1 = \frac{4}{5}$, $p_2 = \frac{2}{4}$, and $\text{BLEU}_2 = 1 \cdot \exp(\frac{1}{2} \cdot \log 0.8 + \frac{1}{2} \cdot \log 0.5) = 0.81$.

METEOR (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014)

- Uses **alignments** between reference and hypothesis to compute scores.

METEOR (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014)

- Uses **alignments** between reference and hypothesis to compute scores.
- Accounts for **different matching criteria**:
  - **Exact**: Match words if their surface forms are identical.

METEOR (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014)

- Uses **alignments** between reference and hypothesis to compute scores.
- Accounts for **different matching criteria**:
    - **Exact**: Match words if their surface forms are identical.
    - **Stem**: Stem words using a language appropriate and match if the stems are identical.

METEOR (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014)

- Uses **alignments** between reference and hypothesis to compute scores.
- Accounts for **different matching criteria**:
  - **Exact**: Match words if their surface forms are identical.
  - **Stem**: Stem words using a language appropriate and match if the stems are identical.
  - **Synonym**: Match words if they share membership in any synonym set according to the WordNet database.

METEOR (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014)

- Uses **alignments** between reference and hypothesis to compute scores.
- Accounts for **different matching criteria**:
  - **Exact**: Match words if their surface forms are identical.
  - **Stem**: Stem words using a language appropriate and match if the stems are identical.
  - **Synonym**: Match words if they share membership in any synonym set according to the WordNet database.
  - **Paraphrase**: Match phrases if they are listed as paraphrases in a language appropriate paraphrase table.

METEOR

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}$$

$$Score = (1 - Pen) \cdot F_{mean}$$

- α is a trained parameter (there are many more, but not shown here for brevity);
- P is **precision**;
- R is **recall**;
- *Pen* is a **fragmentation penalty**.

BEER (Stanojević and Sima'an, 2014)

- Example of a trained metric;
- Developed by a colleague of ours in the ILLC (Miloš Stanojević);
- Core idea: **integrate different features** in a linear model and **train** the metric.

BEER

- Assume a linear model with features $\vec{\phi}$ and weight vector $\vec{\boldsymbol{w}}$:
  - $\text{score}(h,r) = \vec{\boldsymbol{w}} \cdot \vec{\phi}(h,r)$
- There are human judgements that say that a translation h_{good} is better than a translation h_{bad}.

$$
\begin{aligned}
\text{score}(h_{\text{good}}, r) > \text{score}(h_{\text{bad}}, r) &\iff \\
\vec{\boldsymbol{w}} \cdot \vec{\phi}_{\text{good}} > \vec{\boldsymbol{w}} \cdot \vec{\phi}_{\text{bad}} &\iff \\
\vec{\boldsymbol{w}} \cdot \vec{\phi}_{\text{good}} - \vec{\boldsymbol{w}} \cdot \vec{\phi}_{\text{bad}} > 0 &\iff \\
\vec{\boldsymbol{w}}(\vec{\phi}_{\text{good}} - \vec{\phi}_{\text{bad}}) > 0 & \\
\vec{\boldsymbol{w}}(\vec{\phi}_{\text{bad}} - \vec{\phi}_{\text{good}}) < 0 &
\end{aligned}
$$

- This transforms the task from a **ranking task** into a **binary classification task** (positive vs. negative).

WMT Evaluation Shared Task [1]

http://www.statmt.org/wmt16/pdf/W16-2302.pdf

	cs-en		de-en		fi-en		ro-en		ru-en		tr-en	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	6	6	10	10	9	9	7	7	10	10	8	8
MPEDA	**.996**	**.993**	**.956**	.937	**.967**	**.976**	**.938**	**.932**	**.986**	.929	**.972**	**.982**
UoW.ReVal	**.993**	**.986**	**.949**	**.985**	**.958**	**.970**	**.919**	**.957**	**.990**	**.976**	**.977**	**.958**
BEER	**.996**	**.990**	.949	.879	**.964**	**.972**	**.908**	.852	**.986**	.901	**.981**	**.982**
chrF1	.993	.986	.934	.868	**.974**	**.980**	.903	.865	**.984**	.898	**.973**	.961
chrF2	**.992**	.989	.952	.893	**.957**	**.967**	.913	.886	**.985**	.918	.937	.933
chrF3	.991	.989	**.958**	.902	.946	.958	.915	**.892**	**.981**	.923	.918	.917
CharacTer	**.997**	**.995**	**.985**	.929	.921	.927	**.970**	**.883**	.955	**.930**	.799	.827
mtevalNIST	.988	.978	.887	.801	.924	.929	.834	.807	.966	.854	**.952**	**.938**
mtevalBLEU	**.992**	**.989**	.905	.808	.858	.864	.899	.840	.962	.837	.899	.895
mosesCDER	**.995**	**.988**	.927	.827	.846	.860	.925	.800	.968	.855	.836	.826
mosesTER	**.983**	**.969**	.926	.834	.852	.846	.900	.793	.962	.847	.805	.788
wordF2	**.991**	**.985**	.897	.786	.790	.806	.905	.815	.955	.831	.807	.787
wordF3	**.991**	**.985**	.898	.787	.786	.803	.909	.818	.955	.833	.803	.786
wordF1	**.992**	**.984**	.894	.780	.796	.808	.890	.804	.954	.825	.806	.776
mosesPER	.981	.970	.843	.730	.770	.767	.791	.748	**.974**	.887	**.947**	.940
mosesBLEU	**.991**	**.983**	.880	.757	.752	.759	.878	.793	.950	.817	.765	.739
mosesWER	**.982**	**.967**	.926	.822	.773	.768	.895	.762	.958	.837	.680	.651

newstest2016

Table 4: Absolute Pearson correlation of to-English system-level metric scores with human assessment variants: RR = standard WMT relative ranking; DA = direct assessment of translation adequacy.

WMT Evaluation Shared Task [2]

http://www.statmt.org/wmt16/pdf/W16-2302.pdf

	en-cs		en-de		en-fi		en-ro		en-ru		en-tr	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	10		15		13		12		12	12	8	
CharacTER	**.947**	-	**.915**	-	**.933**	-	**.959**	-	**.954**	**.966**	**.930**	-
BEER	**.973**	-	.732	-	.940	-	**.947**	-	.906	.922	**.956**	-
chrF2	.954	-	.725	-	**.974**	-	.828	-	.930	**.955**	**.940**	-
chrF3	.954	-	.745	-	**.974**	-	.818	-	**.936**	**.960**	**.916**	-
mosesCDER	**.968**	-	.779	-	.910	-	**.952**	-	.874	.874	.791	-
chrF1	.955	-	.645	-	.931	-	.858	-	.901	.928	**.938**	-
wordF3	**.964**	-	.768	-	.901	-	**.931**	-	.836	.840	.714	-
wordF2	**.964**	-	.766	-	.899	-	**.933**	-	.836	.840	.715	-
wordF1	**.964**	-	.756	-	.888	-	**.937**	-	.836	.839	.711	-
MPEDA	**.964**	-	.684	-	**.944**	-	.786	-	.856	.866	.860	-
mosesBLEU	**.968**	-	.784	-	.857	-	**.944**	-	.820	.820	.693	-
mtevalBLEU	**.968**	-	.752	-	.868	-	.897	-	.835	.838	.745	-
mtevalNIST	**.975**	-	.625	-	.886	-	.882	-	.890	.897	.788	-
mosesTER	.940	-	.742	-	.863	-	.906	-	.882	.879	.644	-
mosesWER	.935	-	.771	-	.855	-	**.912**	-	.882	.876	.570	-
mosesPER	**.974**	-	.681	-	.700	-	**.944**	-	.857	.854	.641	-
chrF3.2ref	-	-	-	-	**.973**	-	-	-	-	-	-	-
chrF2.2ref	-	-	-	-	**.970**	-	-	-	-	-	-	-
chrF1.2ref	-	-	-	-	.923	-	-	-	-	-	-	-
wordF3.2ref	-	-	-	-	.890	-	-	-	-	-	-	-
wordF2.2ref	-	-	-	-	.887	-	-	-	-	-	-	-
wordF1.2ref	-	-	-	-	.876	-	-	-	-	-	-	-

newstest2016

Table 5: Absolute Pearson correlation of out-of-English system-level metric scores with human assessment variants: RR = standard WMT relative ranking; DA = direct assessment of translation adequacy.

Conclusions

- MT evaluation is important for **system tuning** and assessing **how good a system is**;
- Different MT metrics: BLEU, METEOR, BEER.

Future work:

- Quality estimation (evaluation of MT output without references);
- Statistical significance testing;
- Corpus- versus sentence-level metrics;
- Hopefully we can talk about them some other time... ☺

References I

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Stanojević, M. and Sima'an, K. (2014). Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.