The UPC Submission to the WMT 2012 Shared Task on Quality Estimation for Machine Translation

Daniele Pighin, Meritxell Gonzàlez and Lluís Màrquez - Universitat Politècnica de Catalunya, Barcelona (Spain)

Approach overview

- Extend the 17 baseline features with **39** novel features modeling **dependency** and **sequential** aspects
- All features based on reference (**ref**) and automatic translations (**sys**) for 150K WMT newswire translations
- All features involve **projecting linguistic information** from src to tgt via **automatic alignments**
- **Learning framework:** Support Vector Regression (**SVM-Light**), linear kernel, 5-fold cross-validation

For more informations

Check out the details in the paper, and if you have more questions do not hesitate to send us an email:

{pighin,mgonzalez,lluism}@lsi.upc.edu

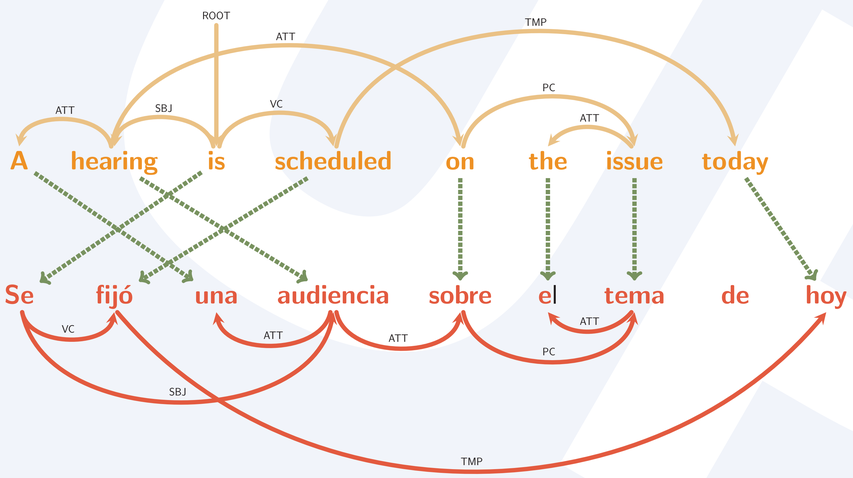
Dependency features - DEP [18]

Assumption: by observing how dependency parses are projected from source to target we can gather clues concerning translation quality that cannot be captured by sequential models

Estimation: count how many times a projected dependency of type D between two words w_1 and w_2 on reference and automatic translations is the same (M_D^+) or different (M_D^-) We estimate three models for word forms (**W**), roots (**R**) and suffixes (**S**)

We encode as features

- **Coverage (COV):** the ratio of projected dependencies observed on training data
- **C$^+$** and **C$^-$**: the number of projected dependencies observed in reference (+) or automatic (-) translations, divided by the total number of projected dependencies, i.e.: $C^+ = \frac{1}{|T|}\sum_{D\in T}\frac{M_D^+}{M_D^+ + M_D^-}$, $C^+ = \frac{1}{|T|}\sum_{D\in T}\frac{M_D^+}{M_D^- + M_D^-}$
- These features encode the likelihood of the projected dependencies to be observed in reference/automatic translations
- The projected edges are sorted, and the above features are estimated on the first (**Q1**), first two (**Q12**), last two (**Q34**) and fourth (**Q4**) quartiles of the distribution



Sequential features - SEQ [21]

Observation: LM-features have the highest correlation with human quality assessments

We estimate **3-gram tgt-side language models** on sequences of:

- Words (**W**), roots (**R**) and suffixes (**S**)
- Stop words on the target side, where non-stop words are replaced by
  - their root (**RStop**) or stem (**SStop**)
  - the POS tag of the aligned source word (**PStop**)
- Chains of stop-words (**Chains**), where adjacent stop words are merged as one token

For each type of sequence, we encode as features

- The perplexity of the LM estimated on **sys**, e.g., **SEQ/sys/PStop**
- The perplexity of the LM estimated on **ref**, e.g., **SEQ/ref/PStop**
- The **ratio** of the two perplexity, e.g., **SEQ/sys-ref/PStop**

Evaluation and discussion

- Marginal improvement on dev data, accuracy loss on the test set

Development set

Feature set	DeltaAvg	MAE
Baseline	0.4664	0.6346
Extended	**0.4694**	**0.6248**

Test set

System	DeltaAvg	MAE
Baseline	**0.55**	**0.69**
Official Evaluation	0.22	0.84
Amended Evaluation	0.51	0.71

- Results (after bug fixing) in line with most other submissions
- Difficult to design global indicators that can do better than the strong baseline features
- Src/Tgt n-gram LMs inherently correlated with post-editing effort
- Learning framework/parametrization much more important than feature-set
- Training data insufficient to estimate features describing complex linguistic interactions
- More refined ways to incorporate rich linguistic features are needed

Feature correlation

Feature	Pearson
BL/4 (src LM)	0.3618
BL/5 (tgt LM)	0.3544
BL/12	0.2823
BL/14	0.2675
BL/2	0.2667
BL/1	0.2620
BL/8	0.2575
BL/6	0.2143
DEP/C$^-$/S	0.2072
BL/10	0.2033
DEP/C$^-$/Q12/S	0.1858
BL/17	0.1824
BL/16	0.1725
DEP/C$^-$/W	0.1584
DEP/C$^-$/R	0.1559
DEP/C$^-$/Q12/R	0.1447
DEP/Coverage/W	0.1419
DEP/C$^-$/Q1/S	0.1413
BL/15	0.1368
DEP/C$^+$/Q4/S	0.1257
DEP/Coverage/R	0.1239
SEQ/ref-sys/PStop	0.1181
SEQ/sys/PStop	0.1173
SEQ/sys-ref/PStop	0.1170
DEP/C$^-$/Q12/W	0.1159
DEP/C$^-$/Q1/R	0.1113
DEP/C$^+$/Q34/S	0.0933
BL/9	0.0889
DEP/C$^+$/Q4/R	0.0749
BL/13	0.0741
DEP/C$^-$/Q1/W	0.0726
DEP/C$^+$/Q4/W	0.0718
DEP/C$^+$/Q34/R	0.0687
BL/3	0.0623
DEP/C$^+$/Q34/W	0.0573
SEQ/sys-ref/W	0.0495
SEQ/sys/W	0.0492
SEQ/ref-sys/W	0.0390
BL/7	0.0351
SEQ/sys/SStop	0.0312
SEQ/sys/RStop	0.0301
SEQ/sys-ref/SStop	0.0291
SEQ/sys-ref/RStop	0.0289
DEP/Coverage/S	0.0286
SEQ/ref-sys/S	0.0232
SEQ/ref-sys/R	0.0205
SEQ/ref-sys/RStop	0.0187
SEQ/sys-ref/R	0.0184
SEQ/sys/R	0.0177
SEQ/ref-sys/Chains	0.0125
SEQ/ref-sys/SStop	0.0104
SEQ/sys/S	0.0053
SEQ/sys-ref/S	0.0051
SEQ/sys/Chains	0.0032
SEQ/sys-ref/Chains	0.0014
BL/11	0.0001