ADAPTIVE QUALITY ESTIMATION FOR MACHINE TRANSLATION AND AUTOMATIC SPEECH RECOGNITION

José G. C. de Souza

Advisors: Matteo Negri, Marco Turchi, Marcello Federico
What is MT Quality Estimation?

- Quality control when there are no references
- Real-time estimations
Applications

• Informing the reader of the target language about whether the translation is reliable.
Applications

- Deciding whether the translation is good enough to be published
- Selecting best MT output out of a pool of MT systems

- Deciding whether the translation needs to be post-edited
  - Computer-assisted translation (CAT) scenario
CAT scenario

- Fuzzy match score for translation memory
- MT suggestions require scores: MT QE
Outline

- Quality Estimation
  - Quality Judgments
  - Quality Indicators

- Current (static) MT QE approaches

- Adaptive approaches
  - Online
  - Multitask
  - Online Multitask
Quality Estimation (QE)

- Supervised learning task
- Quality Judgments (labels)
  - Proxy for correctness and usefulness
- Quality Indicators (features)
- Granularity
  - Word
  - Sentence
  - Document

Source segments → Translated segments

QE Train

QE model

Labels

Train Labels QE model
Quality Judgments

- **Perceived post-editing effort** (Specia, 2011)
  - Two levels of ambiguity

- **Post-editing time** (O’Brien, 2005)
  - High variability

- **Actual Post-editing effort (HTER)** (Tatsumi, 2009)
  - Does not capture cognitive effort
Quality indicators

- Complexity of the source sentence
- Fluency of the translation
- Adequacy of the translation
- MT confidence

QuEst [ACL13a]
Quality indicators

- Complexity of the source sentence:
  - Sentences that are complex at the syntactical, semantic, discursive or pragmatic levels are harder to translate.

- Examples:
  - n-gram language model perplexity
  - average source token length
Quality indicators

- Fluency of the translation
  - Related to grammatical correctness in the target language
  - Example:
    - n-gram language model perplexity
Quality indicators

- **Translation adequacy**
  - Related to the meaning equivalence between source and its translation.

- **Examples:**
  - Ratios of aligned word classes [ACL13b, WMT13, WMT14]
  - Topic-model-based features [MTSummit13]
Quality indicators

- **MT confidence**
  - Related to the difficulty of the MT process
  - **Examples**
    - log-likelihood scores (normalized by source length)
    - average distances between n-best hypotheses [WMT13,14]
Outline

- Quality Estimation
  - Quality Judgments
  - Quality Indicators

- **Current (static) MT QE approaches**

- Adaptive approaches
  - Online
  - Multitask
  - Online Multitask
Problems in current MT QE approaches

- Systems assume ideal conditions:
  - Single MT system, text type and user

- Best setting is task-dependent

- Scarcity of labeled data
MT QE in real conditions

- QE in the CAT scenario typically requires dealing with diverse input:
  - Different genres/types of text/projects
  - Different MT systems
  - Different post-editors

- Here, users + text type + MT system = domain/task
Outline

- Quality Estimation
  - Quality Judgments
  - Quality Indicators

- Current (static) MT QE approaches

- Adaptive approaches
  - Online
  - Multitask
  - Online Multitask
Adaptive QE

- Copes with variability in:
  - Post-editors
  - Text types
  - MT quality
Online QE

Training

Test

Training/Test

Sentence pair

Human feedback

Quality prediction

Adaptive Online QE

Domain$_1$

Domain$_2$

Domain$_2$

Human feedback

Empty Online QE

Quality prediction

Sentence pair

[ACL14]
Online QE

- Explores user corrections to adapt to different post-editing styles and text types

- Online learning for MT QE
  - Passive Aggressive (PA) (Crammer et al., 2006)
  - Online Support Vector Machines (Parrella, 2007)
Results

Online QE improves over batch on very different domains
- Empty more accurate than Adaptive
MT QE across multiple domains

- Online MT QE is not able to deal with several domains at the same time
MT QE across multiple domains

- **Multitask learning** (Caruana 1997)
  - Leverages different domains
  - Knowledge transfer between domains

[Coling14a]
Experimental Setting

- Data: 363 src, tgt and post-edit sentences
  - TED talks transcripts, IT manuals, News-wire texts
  - 181/182 training/test

- Baselines:
  - Single task learning (SVR in-domain)
  - Concatenation of domains (SVR pooling)
  - Frustratingly Easy Domain Adaptation (SVR FEDA) (Daumé, 2007)
MT QE across multiple Domains

Learning curve showing MAE for different amounts training data (95% conf. bands)

- Pooling and FEDA worse than Mean
- Improvements over in-domain models
- RMTL usually requires less in-domain data
What have we learnt so far?

- Online QE methods
  - Continuous learning from user feedback
  - Do not exploit similarities between domains

- Batch multitask learning
  - Models similarities between domains
  - Requires complete re-training
Online Multitask MT QE (PAMTL)

- Combines online learning and multitask learning
  - Based on Passive Aggressive algorithms (Crammer et al. 2006)
    - Epsilon-insensitive loss (regression)

- Identifies task relationships (Saha et al. 2011)
Online Multitask MT QE (PAMTL)

- Interaction matrix is initialized so that tasks are learnt independently.
- After a given number of instances the matrix is updated computing divergences over the task weights.

```
| t₁ | ... | tₙ |
```

Interactive matrix

Model (feature weights)
Experimental Setting (data)

- 1,000 En-Fr tuples of (source, translation, post-edit):
  - TED talks (TED)
  - Educational Material (EM)
  - $(\text{IT}_{\text{LSP1}})$, software manual
  - $(\text{IT}_{\text{LSP2}})$, automotive software manual
  - 700/300 train/test
Experimental Settings (baselines)

- Online learning for QE
  - Passive Aggressive (PA-I)
  - Two usages

- Concatenation of domains (STL\textsubscript{pool}), one for all domains

Single task learning (STL\textsubscript{in}), one per domain
Results (stream of domains)

- Pooling presents very poor performance
- PAMTL outperforms all baselines
- PAMTL MAE with 20% of data ≈ in-domain training with 100% of data
Conclusion

- **Before the work presented here:**
  - Static QE systems serving one domain

- **After the work presented here:**
  - Adaptive QE systems serving diverse domains
Conclusion

- Adaptive approaches that can be used for domain adaptation
  - Single-domain adaptation: online QE
  - Multi-domain adaptation: batch MTL QE
  - Multi-domain with online updates: online MTL QE
Conclusion

- State-of-the-art MT QE features for post-editing time and effort prediction

- Introduction of QE for ASR
  - Adaptive QE for ASR shows improvements over in-domain models for both classification and regression scenarios

- New online multitask algorithm for multi-domain large-scale regression problems
Thank you!
Publications


References


