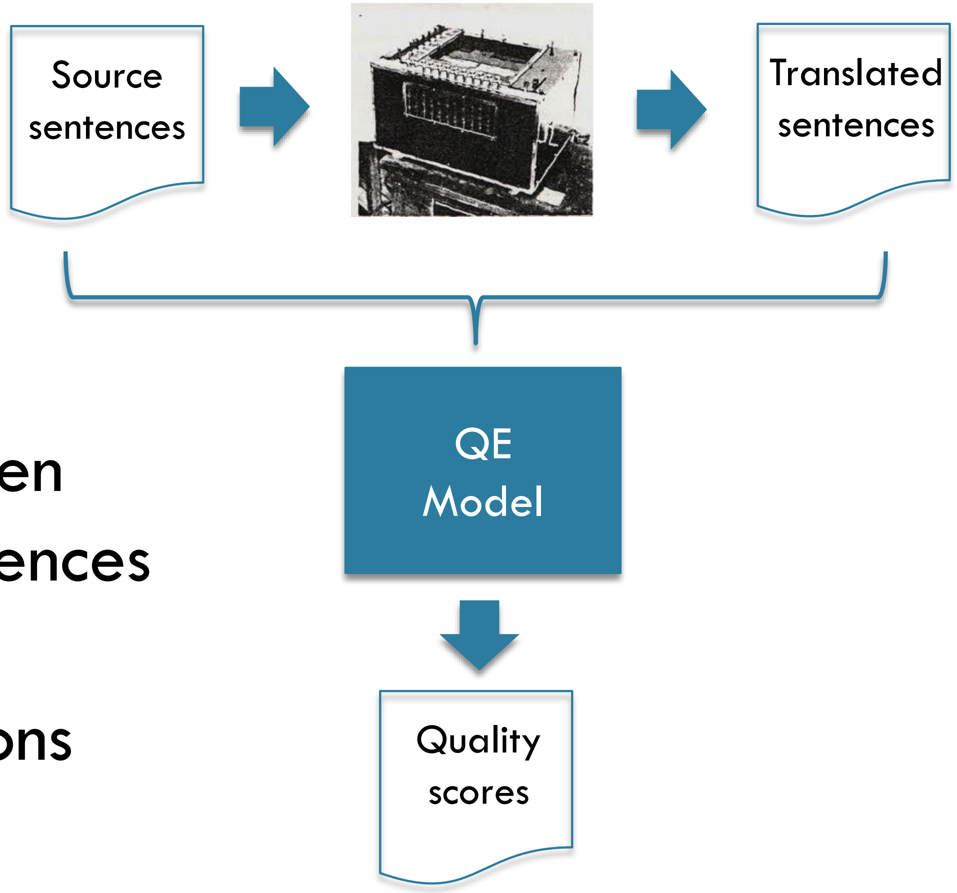
ADAPTIVE QUALITY ESTIMATION FOR MACHINE TRANSLATION AND AUTOMATIC SPEECH RECOGNITION

José G. C. de Souza

Advisors:
Matteo Negri
Marco Turchi
Marcello Federico

EAMT 2017
30/05/2017

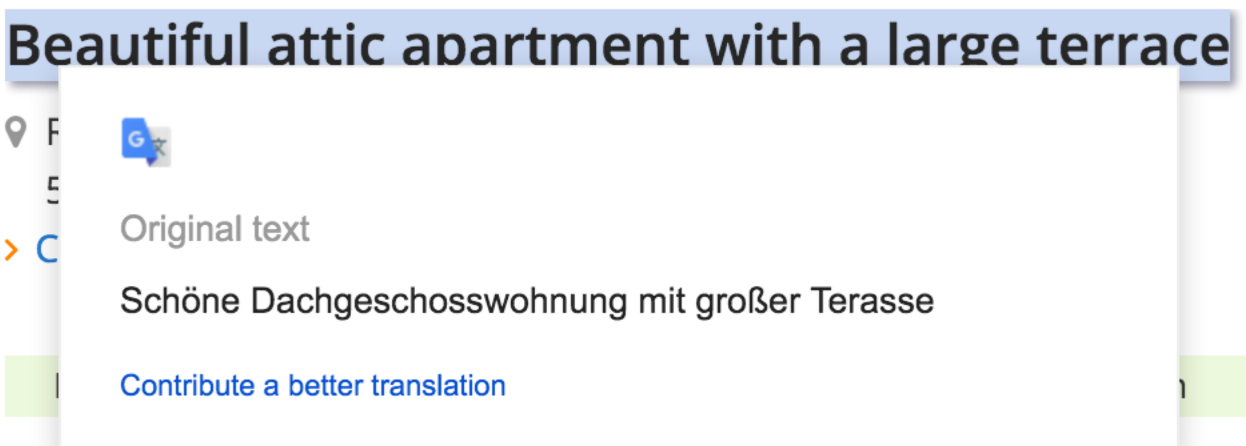
What is MT Quality Estimation?



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

Applications

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

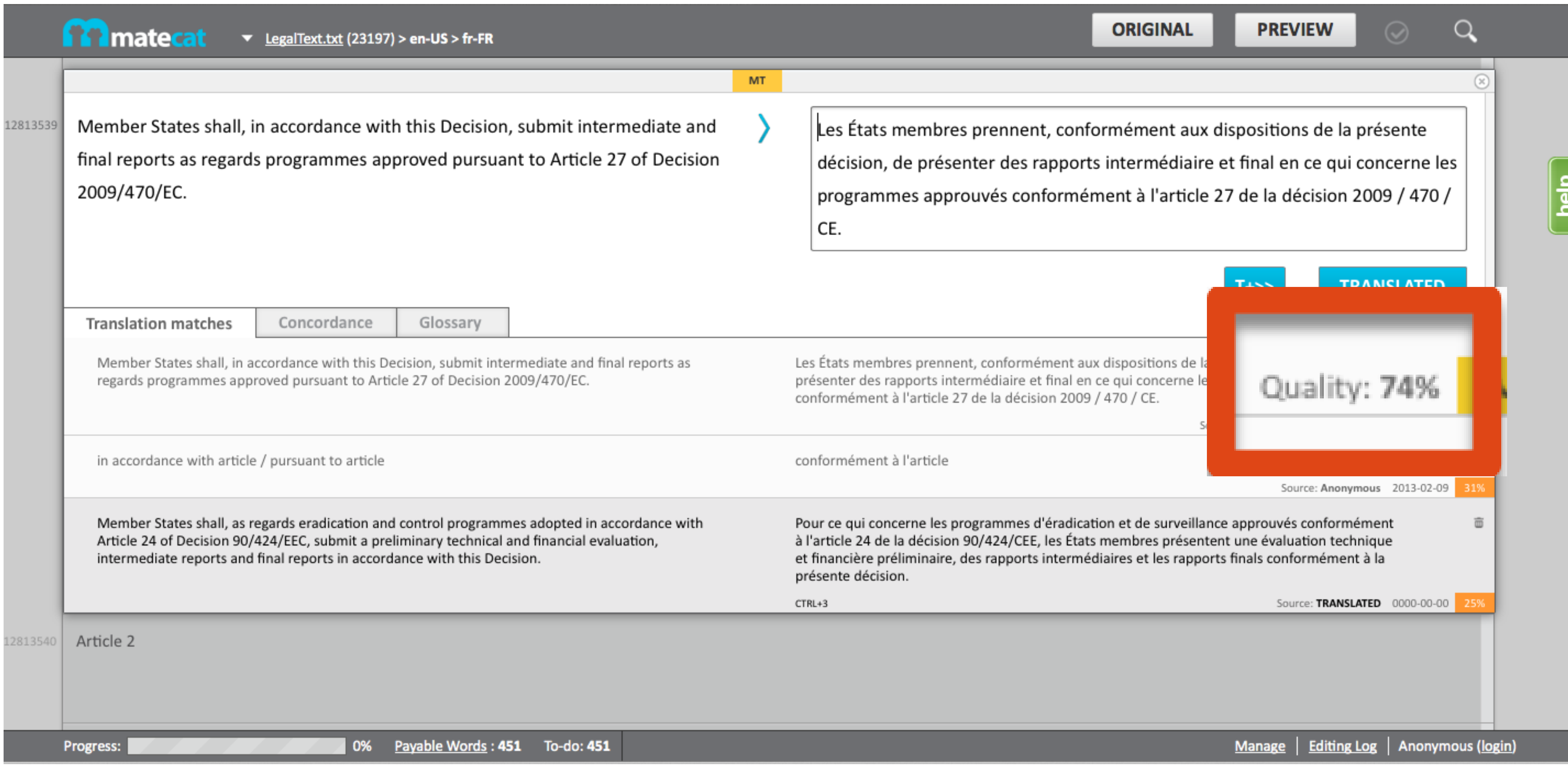


Applications

- 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

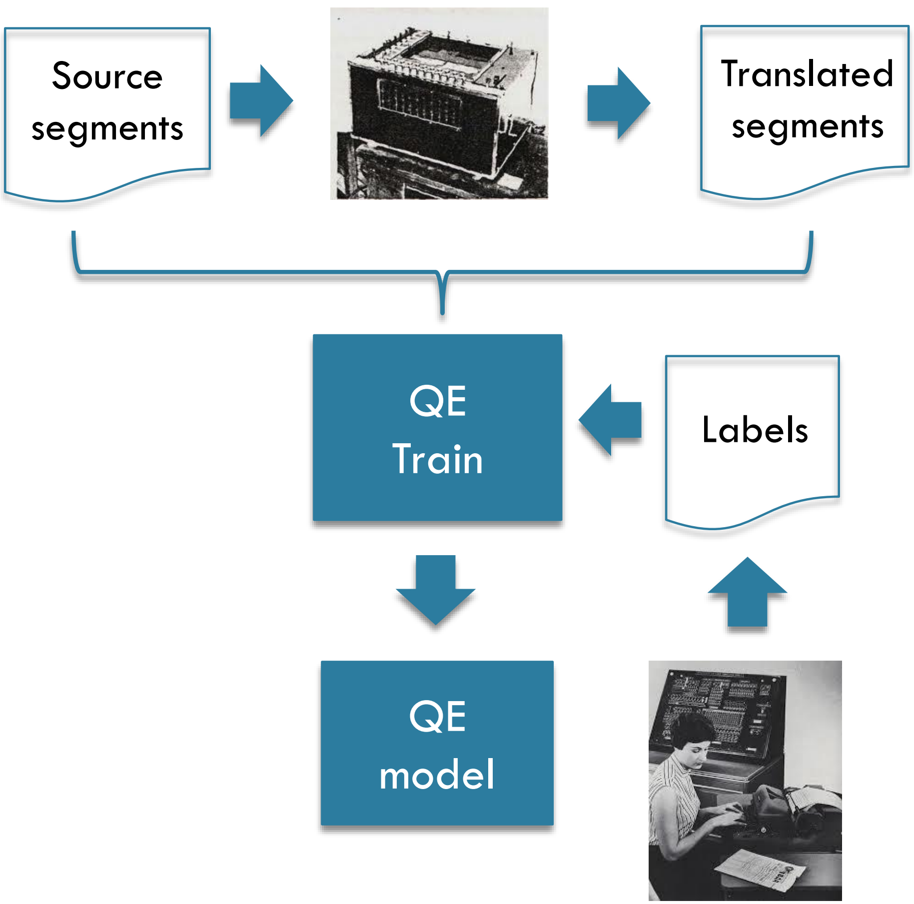
- Fuzzy match score for translation memory
- MT suggestions require scores: MT QE



Outline

- Quality Estimation
  - Quality Judgments
  - Quality Indicators
- Current (static) MT QE approaches
- Adaptive approaches
  - Online
  - Multitask
  - Online Multitask

Quality Estimation (QE)

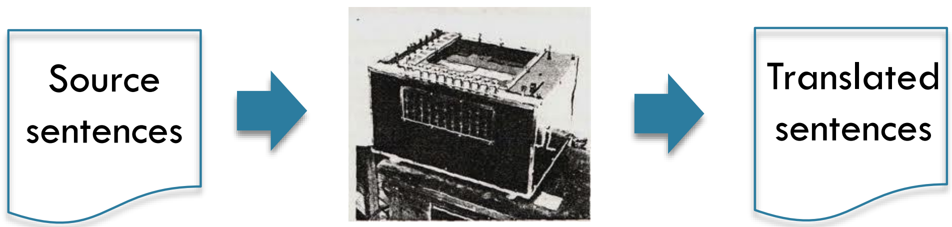


- Supervised learning task
- Quality Judgments (labels)
  - Proxy for correctness and usefulness
- Quality Indicators (features)
- Granularity
  - Word
  - Sentence
  - Document

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

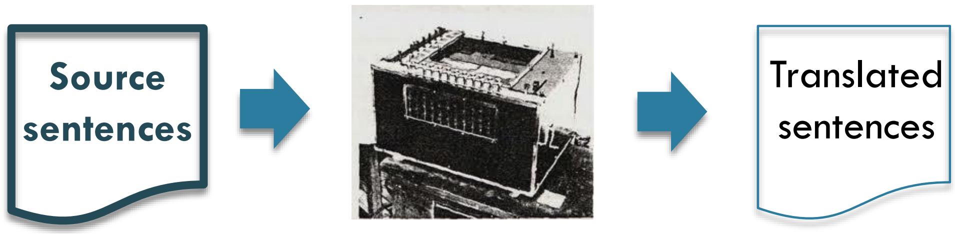
Quality indicators



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

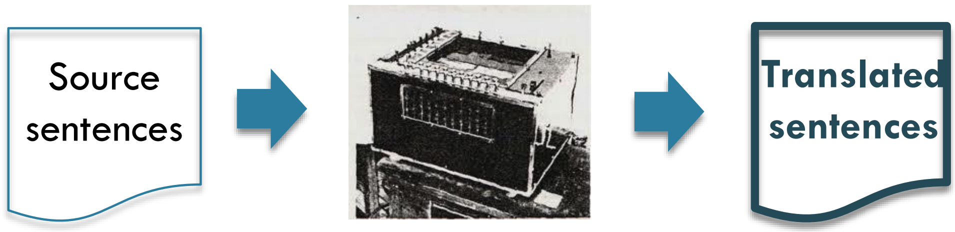
QuEst [ACL13a]

Quality indicators



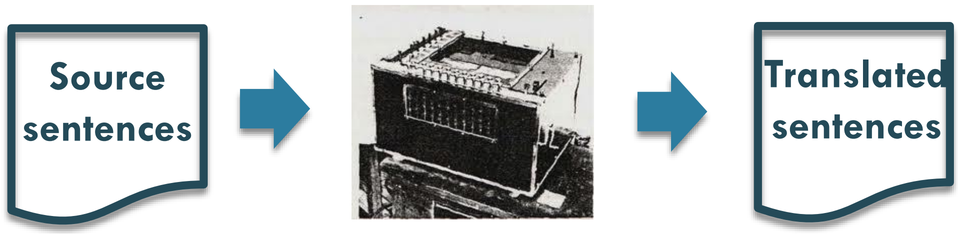
- Complexity of the source sentence;
  - Sentences that are complex at the syntactical, semantic, discursive or pragmatic levels are harder to translate.
  - Examples:
    - n-gram language model perplexity
    - average source token length

Quality indicators



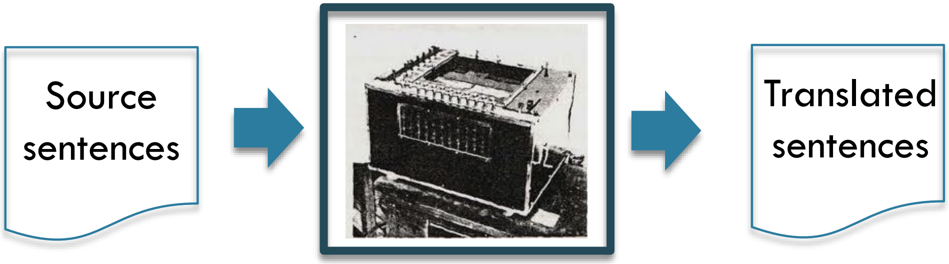
- Fluency of the translation
  - Related to grammatical correctness in the target language
  - Example:
    - n-gram language model perplexity

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 hypothesis [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

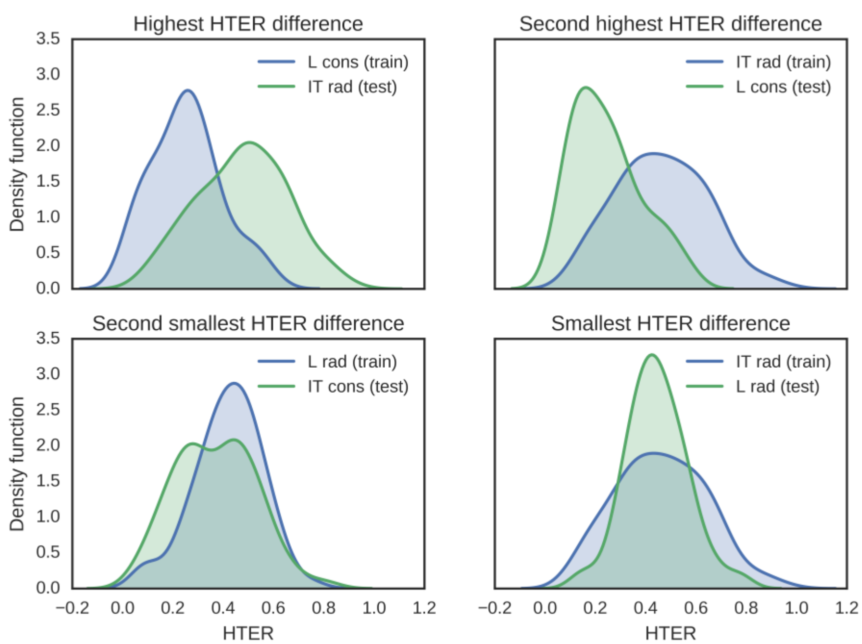
- Here, users + text type + MT system = **domain/task**

Outline

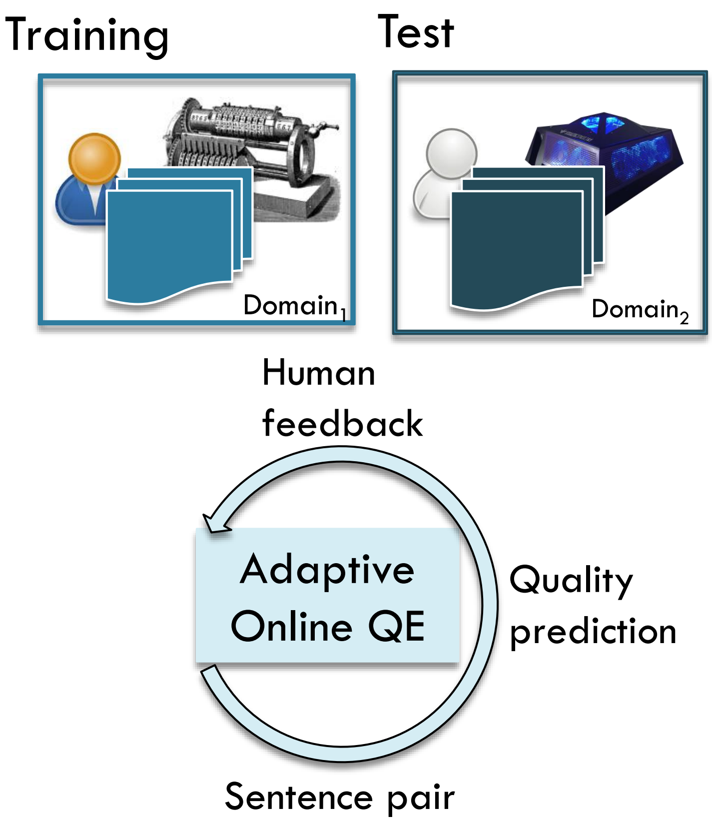
- Quality Estimation
  - Quality Judgments
  - Quality Indicators
- Current (static) MT QE approaches
- **Adaptive approaches**
  - Online
  - Multitask
  - Online Multitask

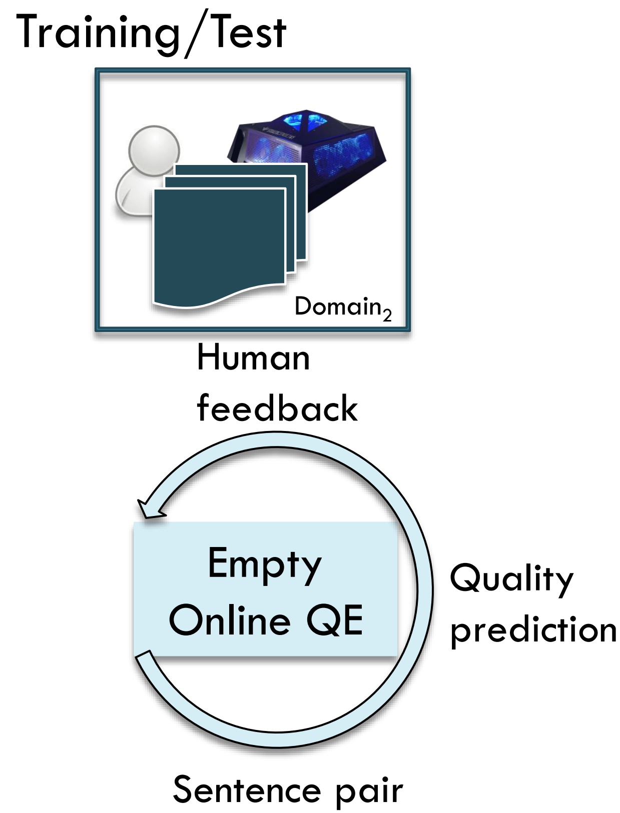
Adaptive QE

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



Online QE





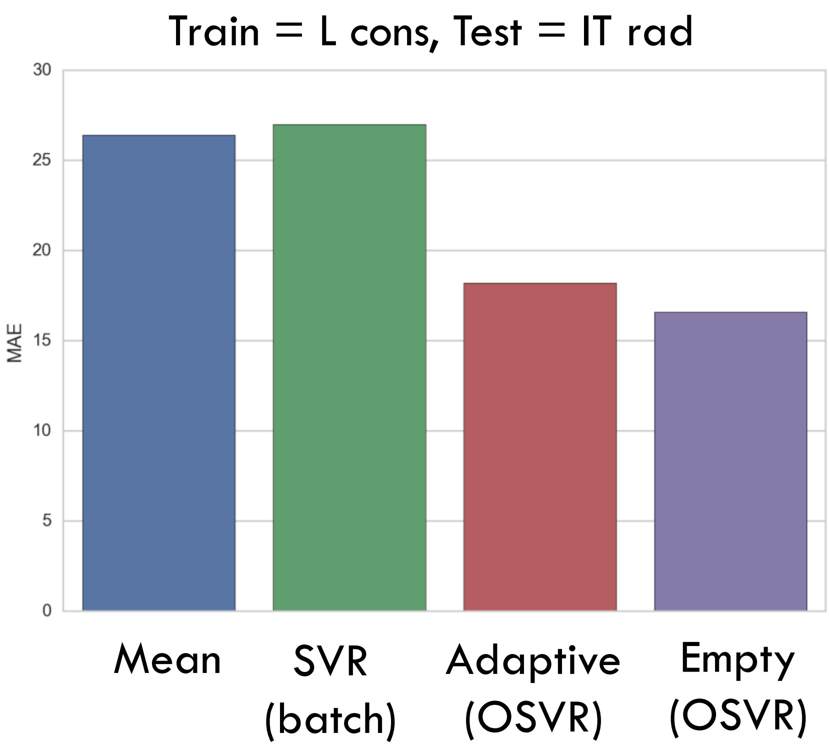
[ACL14]

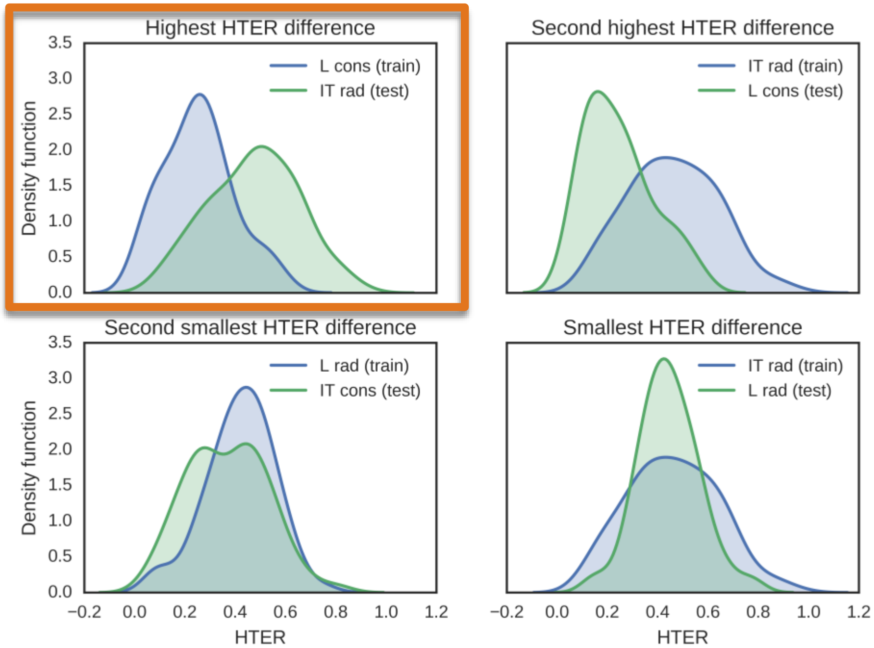
Online QE

- 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

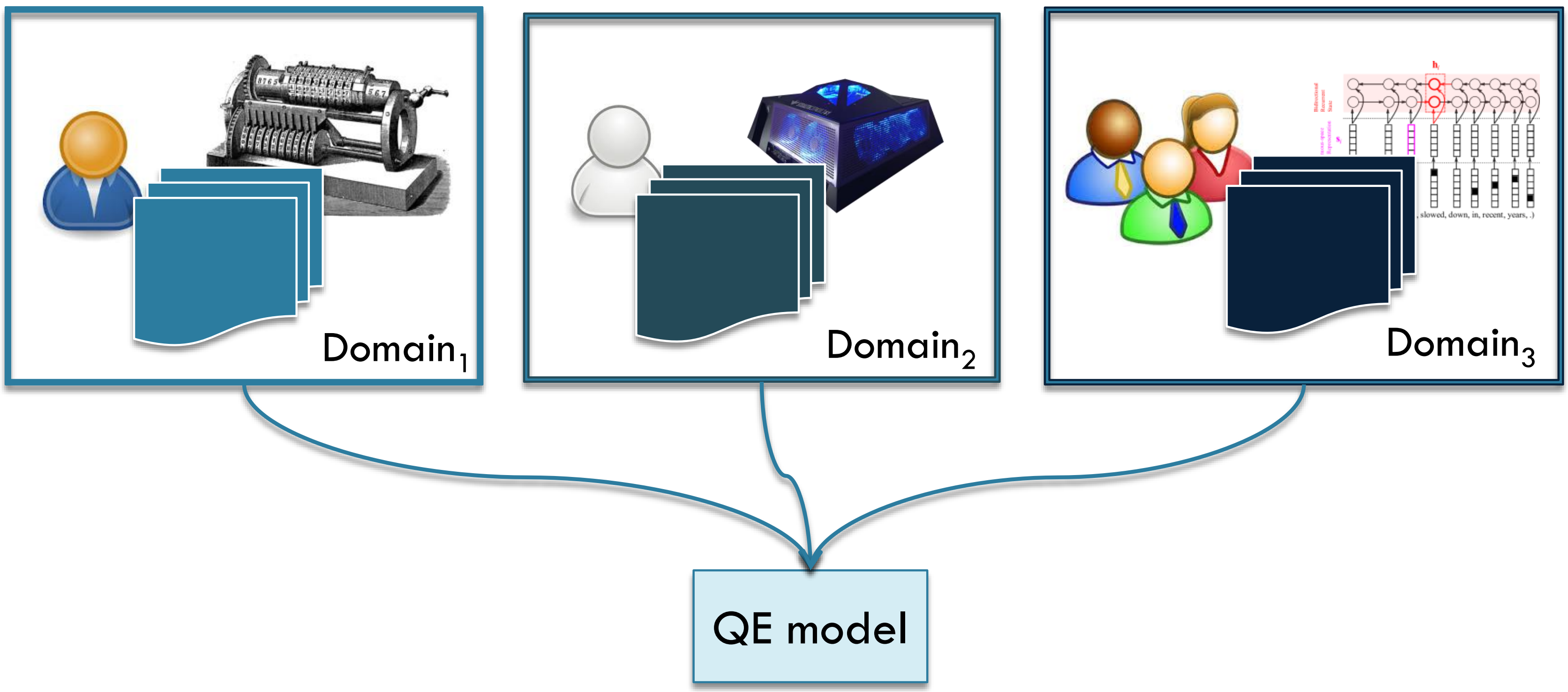




- Online QE improves over batch on very different domains
  - Empty more accurate than Adaptive

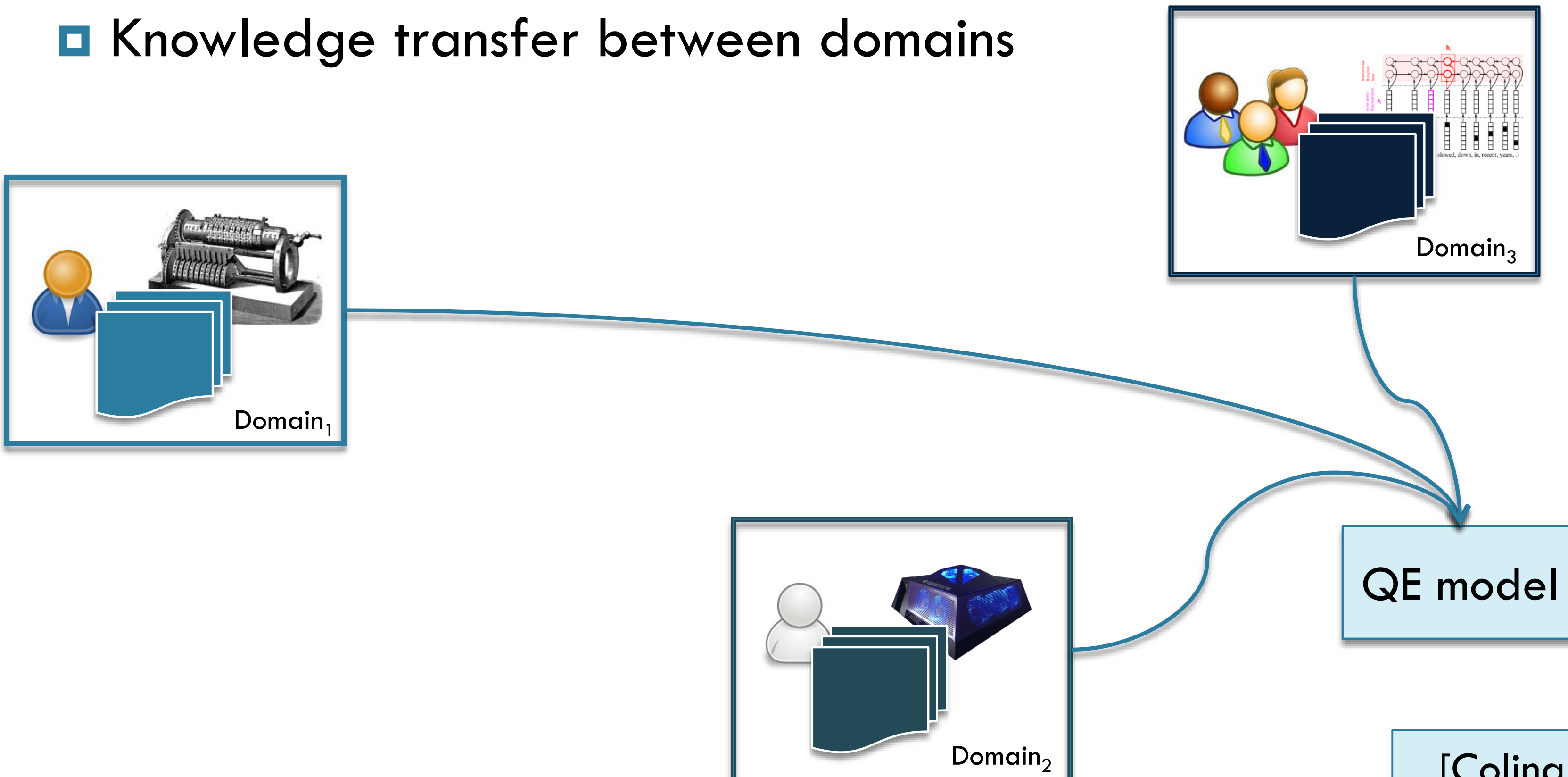
MT QE across multiple domains

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



MT QE across multiple domains

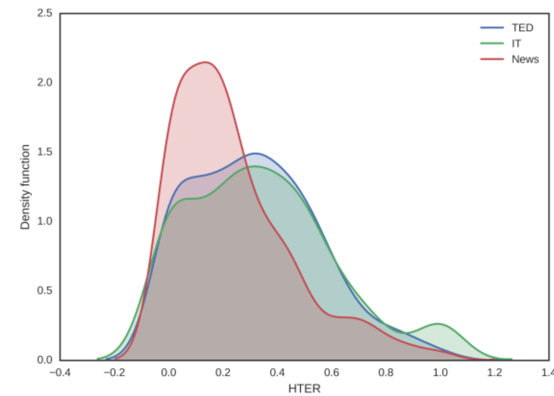
- Multitask learning (Caruana 1997)
  - Leverages different domains
  - Knowledge transfer between domains



[Coling14a]

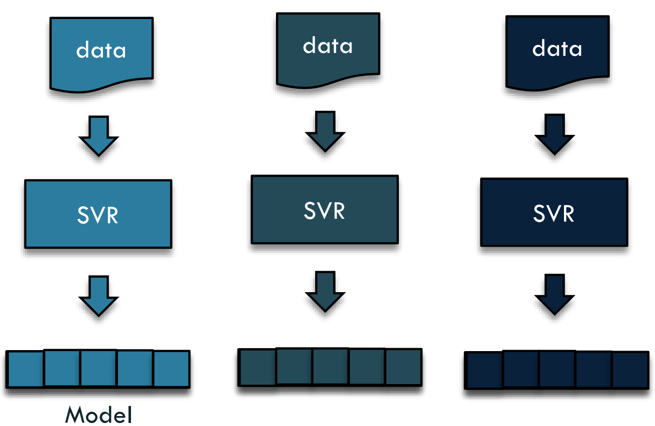
Experimental Setting

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

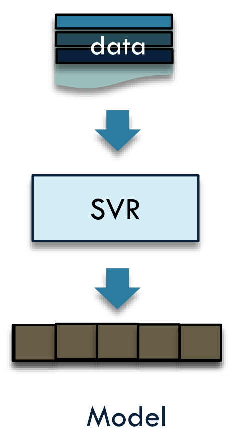


- Baselines:

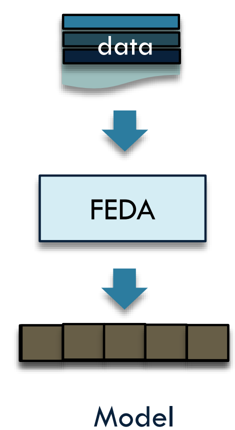
Single task learning (SVR in-domain)



Concatenation of domains (SVR pooling)

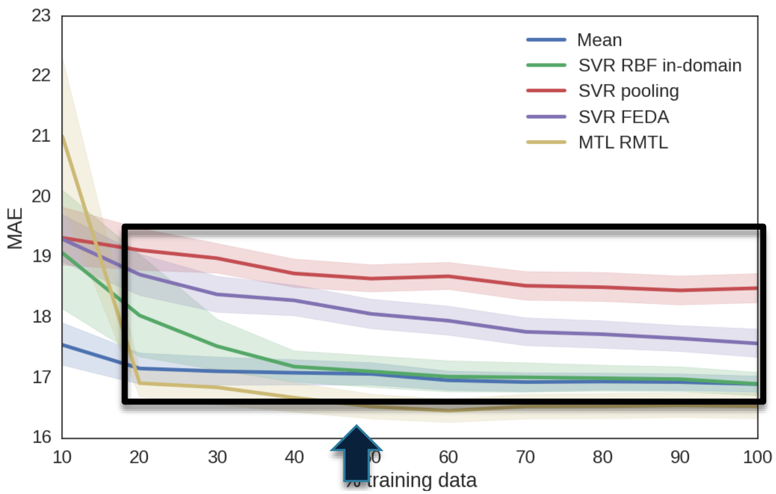


Frustratingly Easy Domain Adaptation (SVR FEDA)(Daumé, 2007)

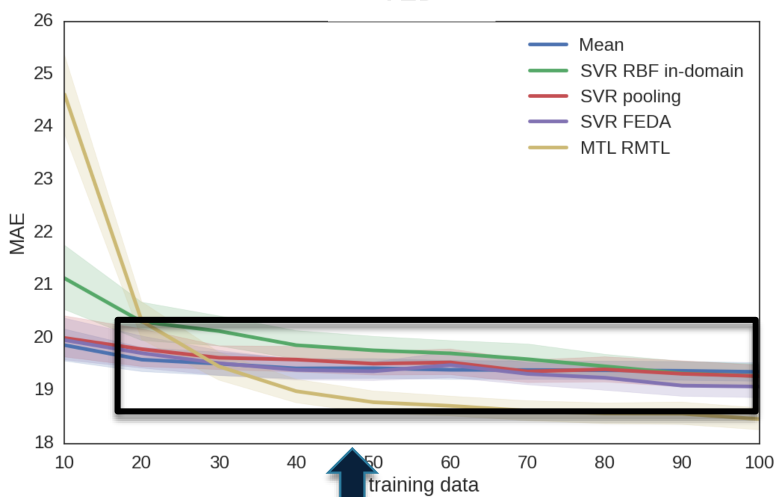


MT QE across multiple Domains

News

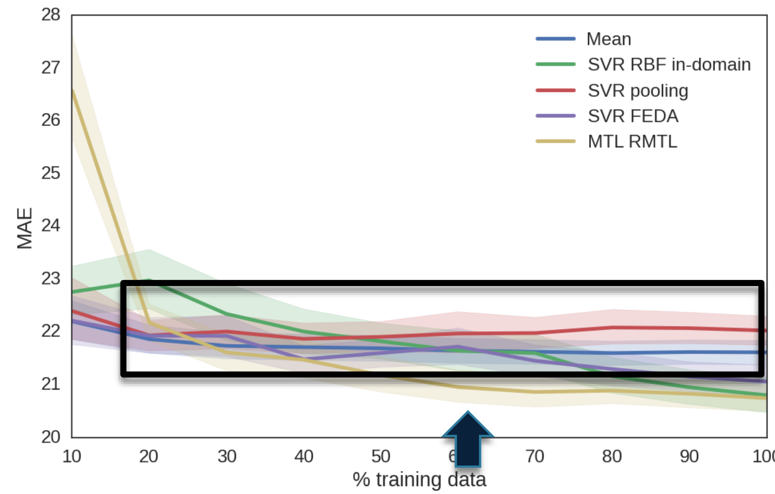


TED



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

IT



- 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

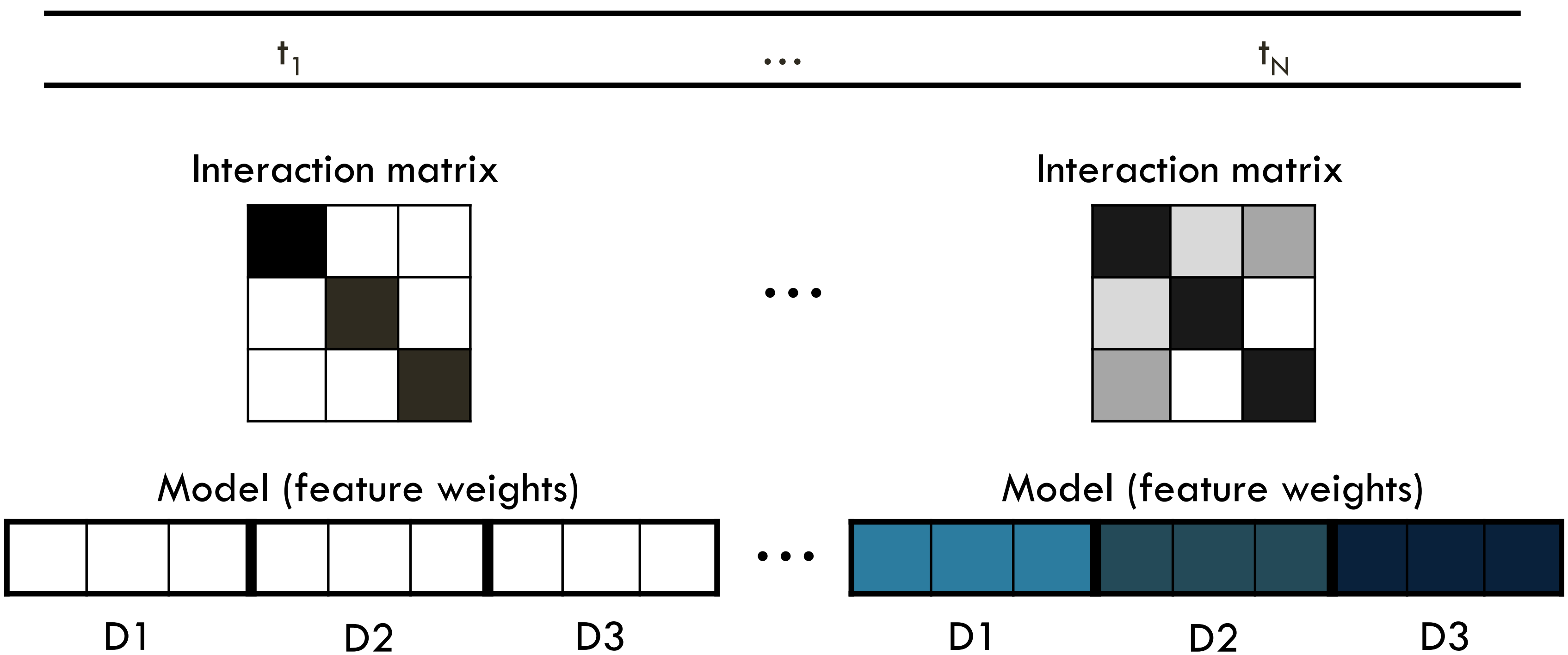
Online Multitask MT QE (PAMTL)

- 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)

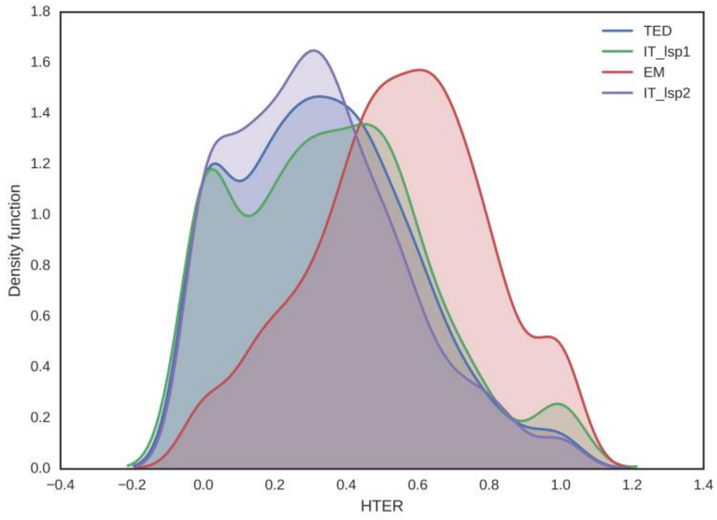
Online Multitask MT QE (PAMTL)

- 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



Experimental Setting (data)

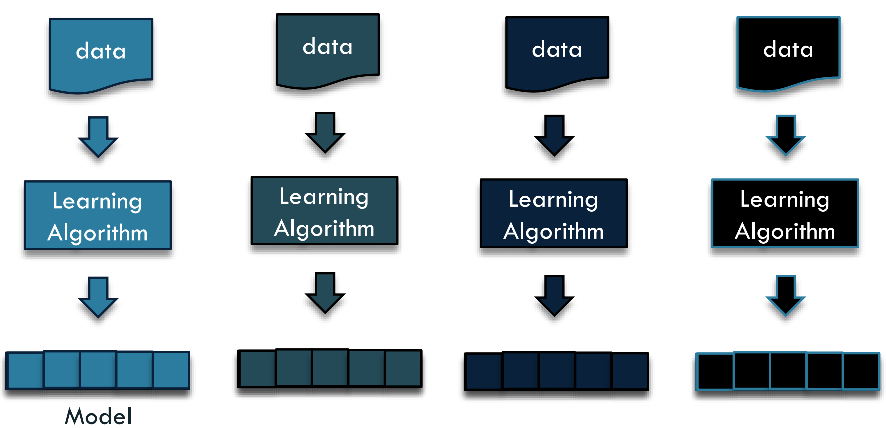
- 1,000 En-Fr tuples of (source, translation, post-edit):
  - TED talks (TED)
  - Educational Material (EM)
  - (IT_{LSP1}), software manual
  - (IT_{LSP2}), automotive software manual
  - 700/300 train/test



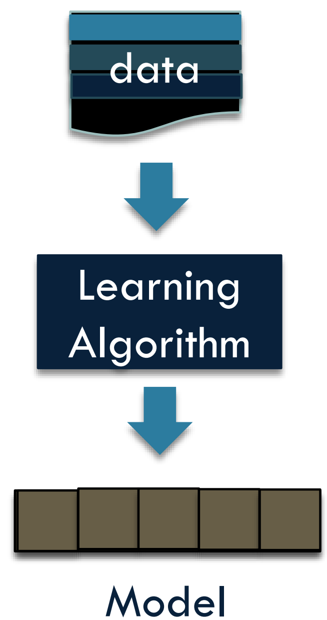
Experimental Settings (baselines)

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

Single task learning (STL_{in}), one per domain

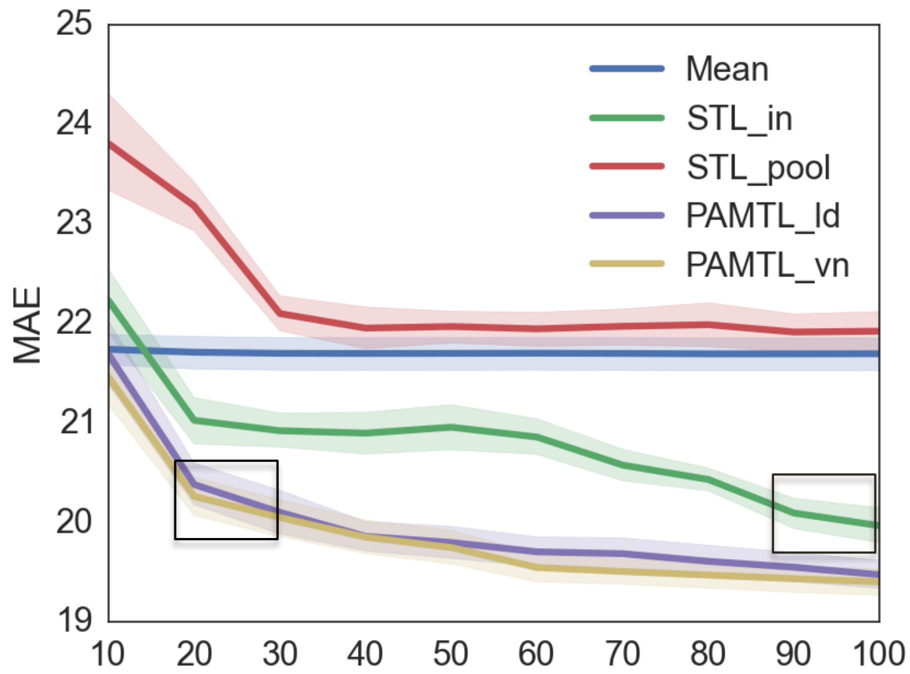


Concatenation of domains (STL_{pool}), one for all domains



Results (stream of domains)

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



- 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

- [WMT13] **José G. C. de Souza**, Christian Buck, Marco Turchi, and Matteo Negri. *FBK-UEdin participation to the WMT13 Quality Estimation shared-task*. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 352–358, 2013
- [ACL13b] **José G. C. de Souza**, Miquel Esplá-Gomis, Marco Turchi, and Matteo Negri. *Exploiting Qualitative Information from Automatic Word Alignment for Cross-lingual NLP Tasks*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 771–776, 2013
- [MTSummit13] Raphael Rubino, **José G. C. de Souza**, and Lucia Specia. *Topic Models for Translation Quality Estimation for Gisting Purposes*. In Machine Translation Summit XIV, pages 295–302, 2013a
- [ACL13a] Lucia Specia, Kashif Shah, **José G. C. de Souza**, and Trevor Cohn. *QuEst–A translation quality estimation framework*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 79–84, 2013

Publications

- [Coling2014a] **José G. C. de Souza**, Marco Turchi, and Matteo Negri. *Machine translation quality estimation across domains*. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 409–420, 2014
- [WMT14] **José G. C. de Souza**, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. *FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task*. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, June 2014a
- [ACL14] Marco Turchi, Antonios Anastasopoulos, **José G. C. de Souza**, and Matteo Negri. *Adaptive Quality Estimation for Machine Translation*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014
- [Coling2014b] Matteo Negri, Marco Turchi, **José G. C. de Souza**, and Falavigna Daniele. *Quality estimation for automatic speech recognition*. In Proceedings of COLING, pages 1813–1823, 2014

Publications

- **José G. C. de Souza**, Marco Turchi, and Matteo Negri. *Towards a combination of online and multitask learning for mt quality estimation: a preliminary study*. In Proceedings of Workshop on Interactive and Adaptive Machine Translation in 2014 (IAMT 2014), 2014b
- [ACL15] **José G. C. de Souza**, Matteo Negri, Elisa Ricci, and Marco Turchi. *Online multitask learning for machine translation quality estimation*. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Inter- national Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL, pages 26–31, 2015
- [NAACL15] **José G. C. de Souza**, Hamed Zamani, Matteo Negri, Marco Turchi, and Daniele Falavigna. *Multitask learning for adaptive quality estimation of automatically transcribed utterances*. Proceedings of NAACL-HLT, Denver, Colorado, pages 714–724, 2015a
- **José G. C. de Souza**, Marcello Federico, and Hassan Sawaf. *MT quality estimation for e-commerce data*. Proceedings of Machine Translation Summit XV, vol. 2: MT Users Track, pages 20–29, 2015b

References

- (Specia, 2011) Lucia Specia. *Exploiting Objective Annotations for Measuring Translation Post-editing Effort*. Proceedings of the 15th Conference of the European Association for Machine Translation, pages 73–80, 2011.
- (O'Brien 2005) Sharon O'Brien. *Methodologies for measuring the correlations between post-editing effort and machine translatability*. Machine Translation, 19(1):37–58, 2005.
- (Tatsumi 2009) Midori Tatsumi. *Correlation between automatic evaluation metric scores, post-editing speed, and some other factors*. The Twelfth Machine Translation Summit (MT-Summit XII), pages 332–339, 2009.

References

- (Crammer et al., 2006) Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. *Online Passive-Aggressive Algorithms*. The Journal of Machine Learning Research, 7:551– 585, 2006.
- (Saha et al. 2011) Avishek Saha, Piyush Rai, Hal Daumé, and Suresh Venkatasubramanian. Online Learning of Multiple Tasks and their Relationships. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, April 2011.
- (Chen et al. 2011)Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group- sparse structures for robust multi-task learning. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11, page 42, New York, New York, USA, 2011. ACM Press.