FONDAZIONE
BRUNO KESSLER

MTM 2014 LAB

Humans in the Loop for MT Improvement: a Hands-on Experience with Manual Error Annotation

L. Bentivogli[1], C. Girardi[1], M. Negri[1], D. Torregrosa Rivero[2], M. Turchi[1]

1. FBK - Fondazione Bruno Kessler

2. University of Alicante

Ninth Machine Translation Marathon,

Trento, Italy, September 9th, 2014

Lab Material

- Annotation guidelines:

  www.statmt.org/mtm14/uploads/Main/Guidelines.pdf

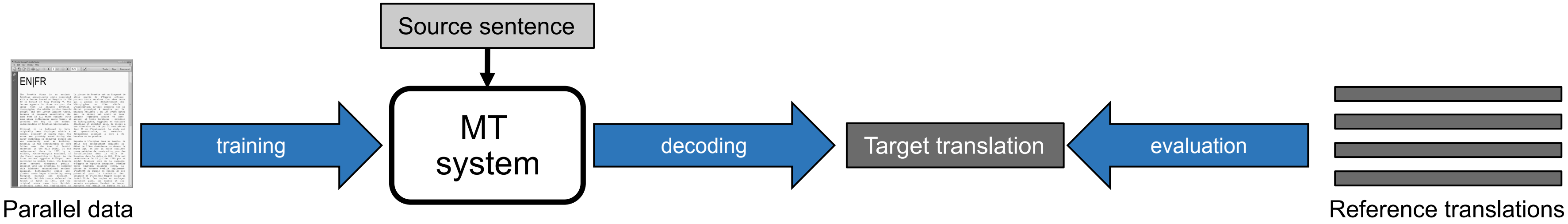
- Annotation results:

  www.statmt.org/mtm14/uploads/Main/Results.pdf

- Lab presentation: next slides, just go on…

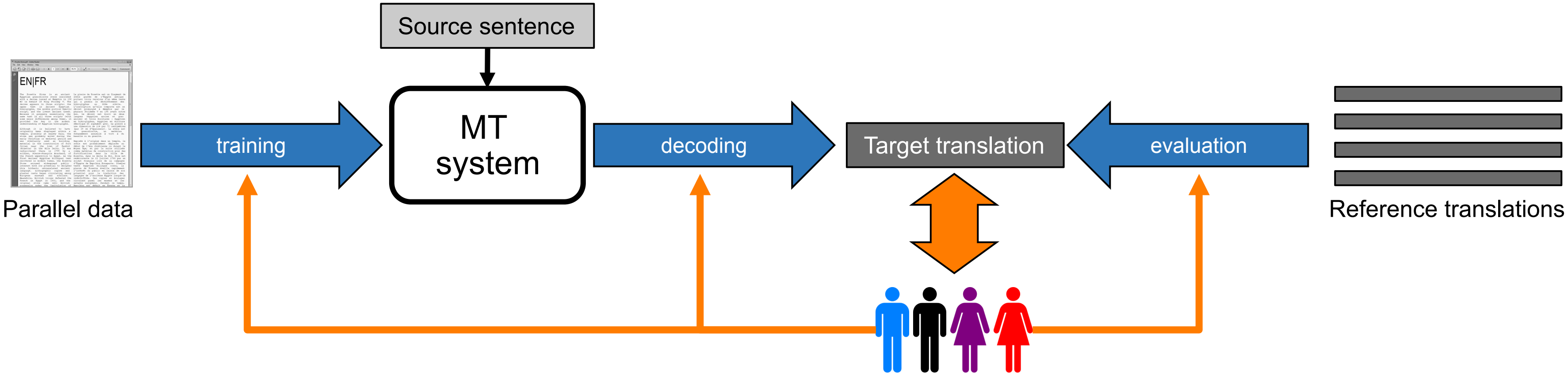
Framework

- SMT typically learns from parallel data and is evaluated by using reference translations



Framework

- SMT typically learns from parallel data and is evaluated by using reference translations



- Further improvements, however, might derive from analyzing different types of human-derived information

Framework

- What types of human feedback can be useful for MT?
- How can we collect useful information from humans?
- What human-annotated corpora are available?
- What tools do we have for the various annotation tasks?
- Is human annotation a difficult task?

Structure of the LAB

- Introduction (15')
  - Framework, uses and types of human annotation, available resources and tools
- MT-EQuAl (15')
  - A Toolkit for Human Assessment of MT output
- Practice: using MT-EQuAl for MT error annotation (40')
- Analysis, discussion and concluding remarks (20')

Possible uses of human annotation/feedback

- **Evaluate** MT output
  - A posteriori, through reference-based automatic metrics (*e.g.* BLEU), fluency/adequacy scores, relative ranking, post-editions (HTER)
  - At run-time (quality estimation), to decide if a given translation is good enough for publishing, inform the readers if they can rely on a translation, filter out bad translations, etc.
- **Improve** MT output production
  - By identifying systems' weaknesses, improving alignment, dynamically modifying phrase tables and language models, etc.
- **Correct** MT output (automatic post-editing)
  - By identifying and correcting recurring errors

Types of human annotation/feedback

- Post-editions
  - Revision of automatic translations
- Quality judgments
  - Scoring (*e.g.* in a 1-to-5 interval) / Relative ranking
- Error annotation
  - Marking MT errors with respect to a given taxonomy

Types of human annotation/feedback

- Post-editions
  - Revision of automatic translations
  - **Natural task**, a by-product of the professional translation workflow
- Quality judgments
  - Scoring (*e.g.* in a 1-to-5 interval) / Relative ranking
  - **Less natural task**, relatively cheap
- Error annotation
  - Marking MT errors with respect to a given taxonomy
  - **Even less natural task**, costly

An additional common problem

All these types of annotation are inherently subjective!

- The same translation can be corrected in different ways [SPE11a]
- Different humans might have different quality standards [COH13,TUR13]
- They might prioritize different errors [LOM14]
- They might produce different rankings [CCB07,CCB08]

- The agreement is often low but…
- …the more the better!
  - For the different tasks, collecting data that account for the variety of human attitudes becomes crucial

Available resources: post-editions

- Few freely available datasets, for few language pairs
  - EN-ES [CCB12,BOJ13,BOJ14], EN-FR [WIS14], FR-EN [WIS13,POT12], EN-IT [TUR14a]
- Typically in the form of [*source, target, reference, post-edition*]
- Sometimes also *HTER scores* and *post-editing time* are provided
- 800 < Size < 10,000 instances
- Used to:
  - Train Quality Estimation components [BOJ14,DES13,DES14,TUR14b]
  - Evaluate and improve SMT systems [POT11,BER13,LOG14,DEN14]
  - Develop automatic post-editing tools [SIM07,BEC13]

Available resources: quality judgments

- Few datasets, not always freely available, for few language pairs
  - EN-ES [CCB08,CCB12,SPE10,TUR14a], EN-AR [SPE11b], FR-EN [SPE09,TUR14a], EN-IT [TUR14a], EN-RU [SPE09]
- Typically in the form of [*source, target, reference, judgment*]
  - Binary “good”/“bad” judgments indicating overall quality
  - Scores based on n-point Likert scales, indicating overall quality/adequacy/fluency
- 700 < Size < 16,000 instances
- Used to:
  - Train Quality Estimation components [MEH12,SPE11b]
  - Evaluate MT systems [GRA14] and automatic metrics [CCB06]

Available resources: error annotation

- Few freely available datasets, for few language pairs
  - EN-CZ, FR-DE, DE-EN, EN-SB [FIS12], EN-FR [WIS14], EN-PT [COS14]
- 60< Size < 2000 instances
- Used to:
  - Identify system's weaknesses [VIL06][STY12][CON10]
  - Train/evaluate error identification tools [POP11] [ZEM11] [BERK12]
  - Train/evaluate error correction tools [SIM07][PAR12][ROS12]
  - …

(Some) available manual annotation tools

- Appraise [FED12]: quality rating/ranking, post-editing, error annotation, web-based
- BLAST [STY11]: error annotation, stand-alone
- PET [AZI12]: post-editing, error annotation (sentence-level), stand-alone
- Translate5 [TRA5]: post-editing, error annotation, web-based
- COSTA [CHA13]: quality rating/ranking, error annotation, stand-alone
- MT-EQuAL [BEN14] quality rating/ranking, error annotation, word-alignment, web-based
- ...

A closer look at MT-EQuAl

- three different tasks in an integrated environment
  - annotation of translation errors
  - translation quality rating
  - word alignment
- web-based, multi-user
- project management functions, configurable tasks
- open source, available on GitHub under Apache 2.0 license

http://mtequal.fbk.eu

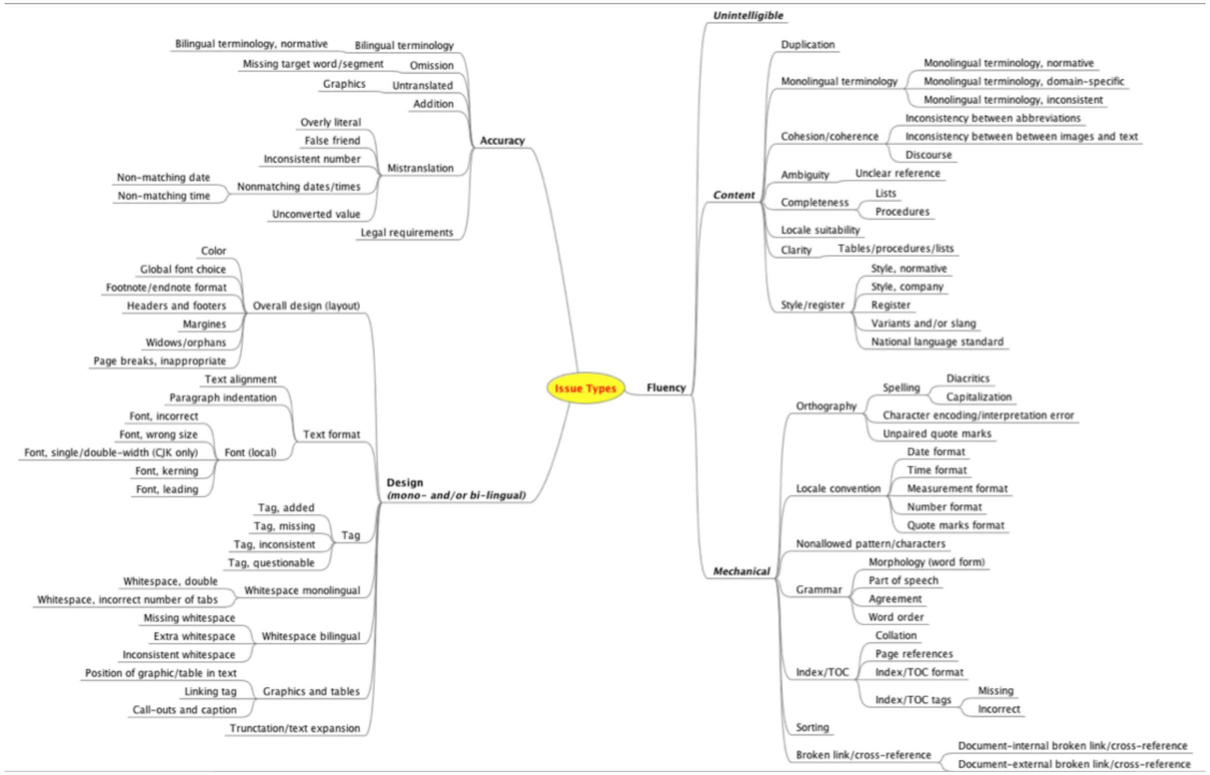
Exercise

Using MT-EQuAl for MT error annotation

Issues in MT error annotation

- Define a reference error **taxonomy**

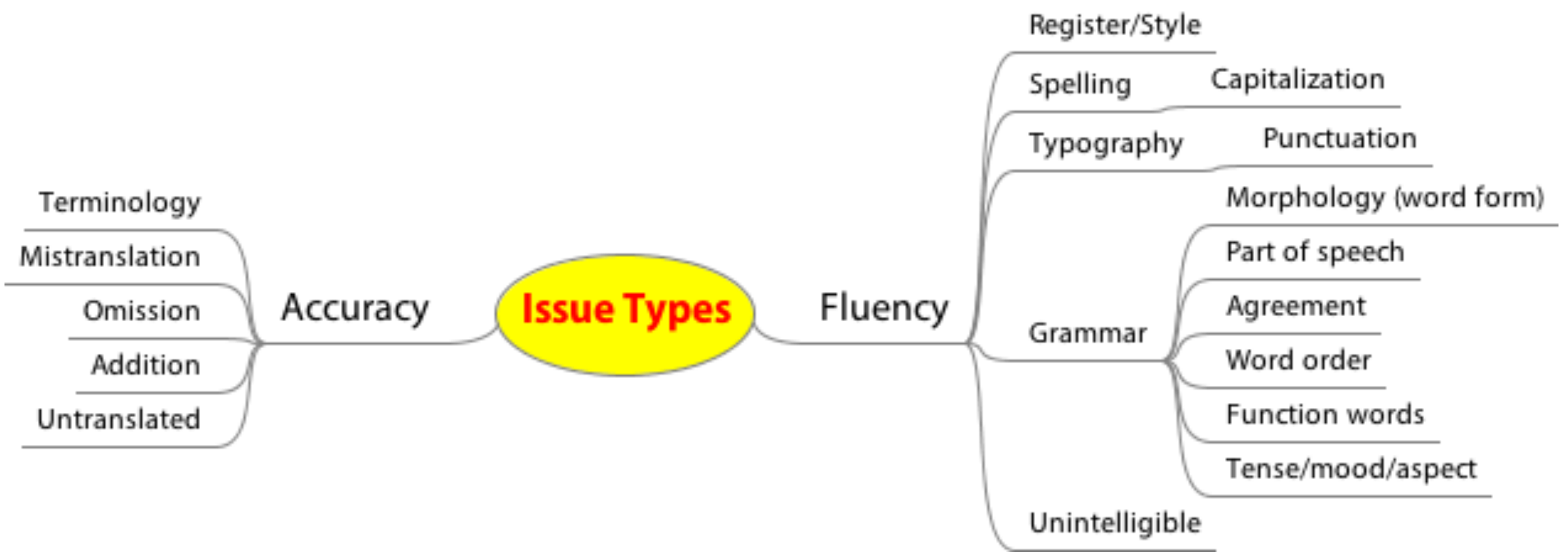
[QTLP]



Issues in MT error annotation

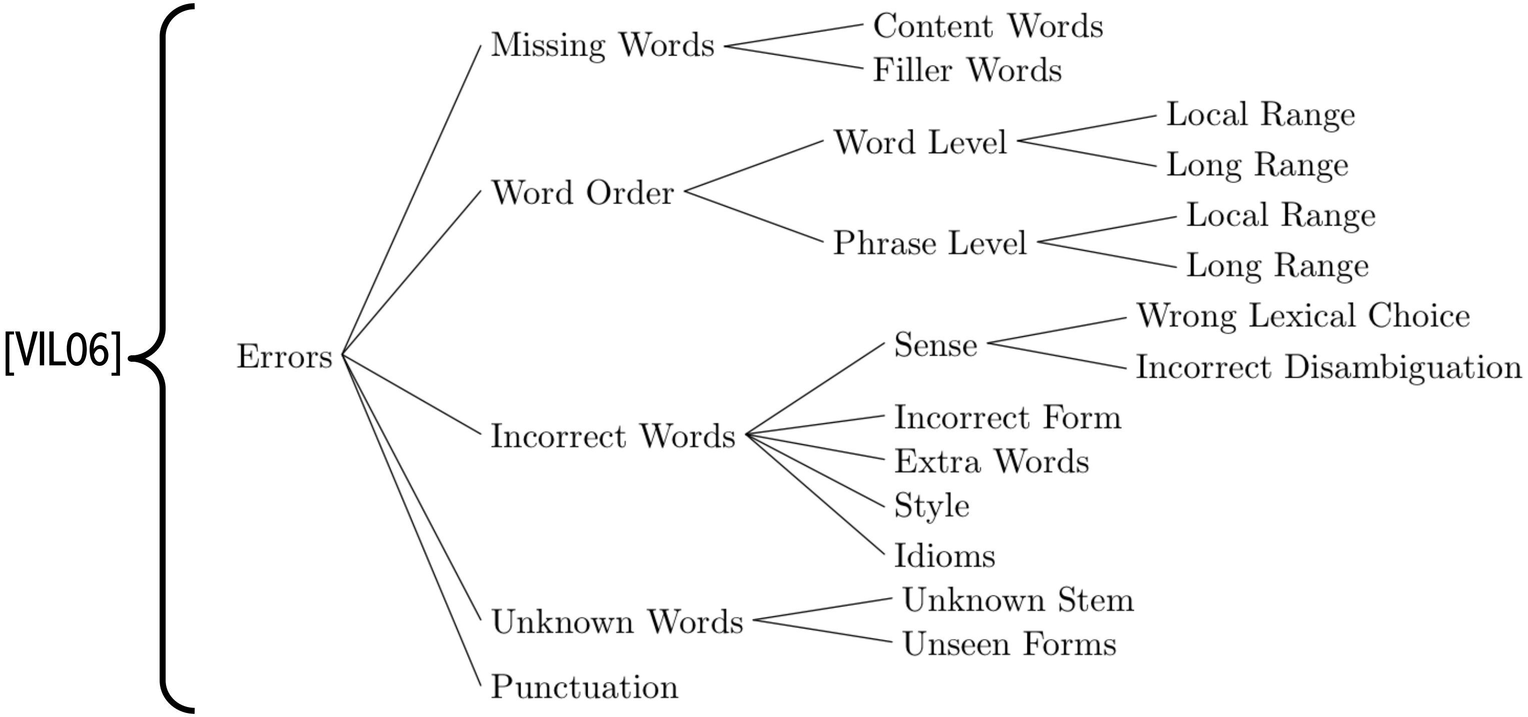
- Define a reference error **taxonomy**

[BOJ14]



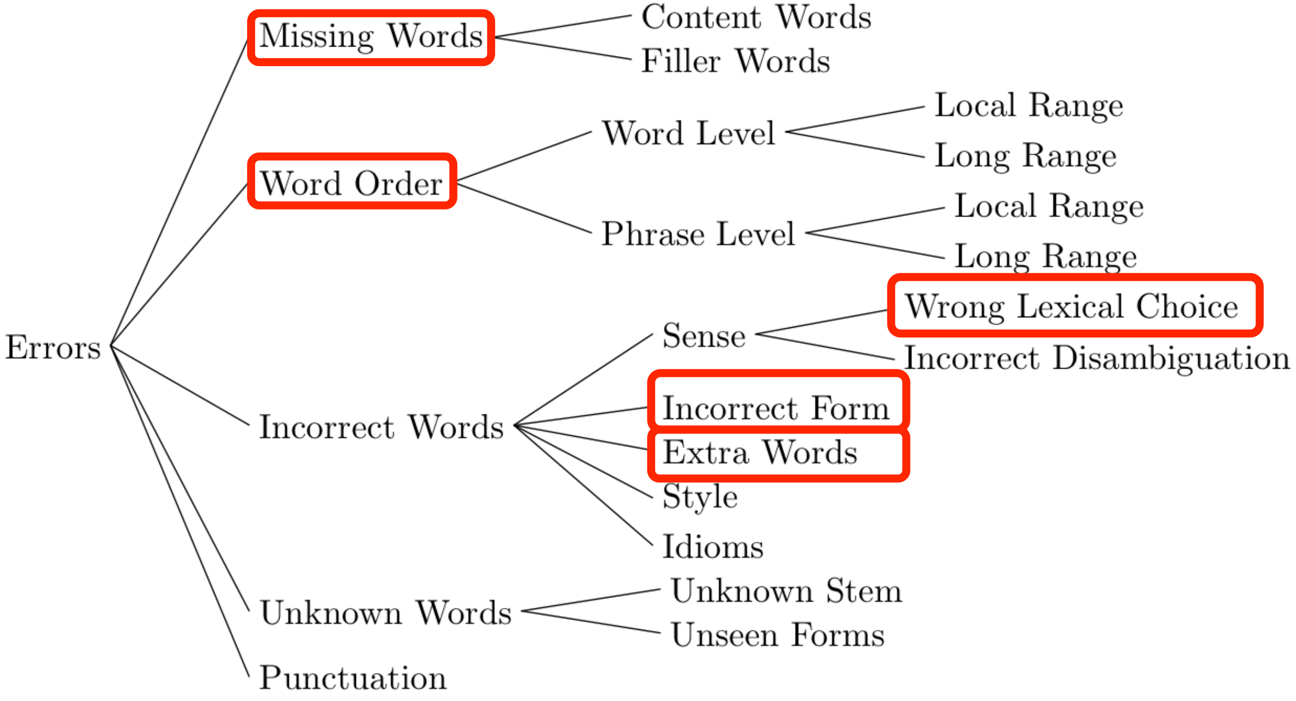
Issues in MT error annotation

- Define a reference error **taxonomy**



Issues in MT error annotation

- Set a **granularity** for the annotation



Issues in MT error annotation

- Annotation based only on the source text or guided by one or more references/post-editions?
- What to annotate?
  - Individual words OR phrases?
  - One OR multiple errors per word?
  - Only the hypothesis OR hypothesis and reference?
- Develop guidelines, train annotators
- …annotate ☺

References

- [AZI12] Aziz, W. Sousa, S. C. M. and Specia, L. (2012). PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey, pp. 3982-3987
URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf

- [BEC13] Bechara, H. (2014) *Statistical post-editing and quality estimation for machine translation systems.* Master of Science thesis, Dublin City University.
URL http://doras.dcu.ie/19751/1/HannaThesis_-_final_submitted.pdf

- [BERK12] Berka, J., Bojar, O., Fishel, M., Popović, M. and Zeman, D. (2012). Automatic MT Error Analysis: Hjerson Helping Addicter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey, pp. 2158-2163.
URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/336_Paper.pdf

- [BER13] Bertoldi, N., Cettolo, M. and Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the XIV Machine Translation Summit.* Nice, France, pp. 35–42
URL http://www.matecat.com/wp-content/uploads/2013/09/mt-summit-2013-bertoldi-et-al.pdf

- [BOJ13] Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R. and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT-2013)*. Sofia, Bulgaria, pp. 1-44.
URL www.statmt.org/wmt13/pdf/WMT01.pdf

- [BOJ14] Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L. and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation,* Baltimore, Maryland*, pp. 12–58.*
URL www.aclweb.org/anthology/W/W14/W14-3302.pdf[CCB07] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 136-158.
URL http://www.statmt.org/wmt07/pdf/WMT18.pdf

- [CCB08] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation.* Columbus, Ohio, pp. 70-106.
URL http://aclweb.org/anthology/W08-0309

References

- [CCB06] Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).* Trento, Italy, *pp. 249–256.*
  URL http://www.aclweb.org/anthology/E/E06/E06-1032.pdf
- [CCB12] Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R. and Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the ACL Seventh Workshop on Statistical Machine Translation (WMT-2012)*. Montreal, Canada, pp. 10-51.
  URL http://www.aclweb.org/anthology/W12-3102
- [CHA13] Chatzitheodorou, K. and Chatzistamatis, S. (2013). COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. The Prague Bulletin of Mathematical Linguistics No. 100, 2013, pp. 83–89.
  URL https://ufal.mff.cuni.cz/pbml/100/art-chatzitheodorou-chatzistamatis.pdf
- [COH13] Cohn, T. and Specia, L. (2013). Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Sofia, Bulgaria, pp. 32-42.
  URL http://www.aclweb.org/anthology/P13-1004
- [CON10] Condon, S., Parvaz, D., Aberdeen, J., Doran, C., Freeman, A. and Awad, M. (2010). Evaluation of Machine Translation Errors in English and Iraqi Arabic. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta, pp. 729-735.
  URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/106_Paper.pdf
- [COS14] Costa, A., Luís, T. and Coheur, L. (2014). Translation Errors from English to Portuguese: an Annotated Corpus. . In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland.
  URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/199_Paper.pdf
- [DEN14] Denkowski, M., Dyer, C. and Lavie, A. (2014). Learning from Post-Editing: Online Model Adaptation for Statistical Machine Translation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, pp. 395-404.
  URL http://www.aclweb.org/anthology/E/E14/E14-1042.pdf
- [DES13] de Souza, J.G.C., Esplà-Gomis, M., Turchi, M. and Negri, M. (2013). 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 (Volume 2: Short Papers)*. Sofia, Bulgaria, pp. 771-776.
  URL http://aclweb.org/anthology/P/P13/P13-2135.pdf

References

- [DES14] de Souza, J.G.C., Turchi, M. and Negri, M. (2014). Predicting Machine Translation Quality Estimation Across Domains. In *Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014)*. Dublin, Ireland, pp. 409-420.
  URL http://www.aclweb.org/anthology/C/C14/C14-1040.pdf
- [FED12] Federmann, C. (2012). Appraise: An open-source toolkit for manual evaluation of machine translation output. The Prague Bulletin of Mathematical Linguistics (PBML), 98:25–35.
  URL https://ufal.mff.cuni.cz/pbml/98/art-federmann.pdf
- [FIS12] Fishel, M., Bojar, O. and Popovic, M. (2012). Terra: a Collection of Translation Error-Annotated Corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey, pp. 7-14.
  URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/481_Paper.pdf
- [GIR14] Girardi, C., Bentivogli, L., Farajian, M.A. and Federico, M. (2014). MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin, Ireland, pp. 120-123.
  URL http://www.aclweb.org/anthology/C14-2026
- [GRA14] Graham, Y., Baldwin, T.J., Moffat, A. and Zobel,, J. (2014). Is Machine Translation Getting Better over Time?. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* Gothenburg, Sweden, pp. 443-451.
  URL http://www.aclweb.org/anthology/E/E14/E14-1047.pdf
- [LOG14] Logacheva, V. and Specia, L. (2014). A Quality-based Active Sample Selection Strategy for Statistical Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland.
  URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/658_Paper.pdf
- [LOM14] Lommel, A., Popović, M. and Burchardt, A. (2014). Assessing Inter-Annotator Agreement for Translation Error Annotation. In *Proceedings of the LREC MTE Workshop on Automatic and Manual Metrics for Operational Translation Evaluation.* Reykjavik, Iceland.
  URL http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=LREC-Lommel-Burchardt-Popovic.pdf&file_id=uploads_2257
- [PAR12] Parton, K., Habash, N., McKeown, K., Iglesias, G. and de Gispert, A. (2012) Can automatic post-editing make MT more meaningful? In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT).* Trento, Italy, pp 111-118
  URL http://www.mt-archive.info/EAMT-2012-Parton.pdf
- [POP11] Popović, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. The Prague Bulletin of Mathematical Linguistics, 96:59–68.
  URL https://ufal.mff.cuni.cz/pbml/96/art-popovic.pdf

References

- [POT11] Potet, M., Esperança-rodier, E., Besacier, L. and Blanchon, H.(2011). Preliminary Experiments on Using Users' Post-Editions to Enhance a SMT System. In *Proceedings of the European Association for Machine Translation (EAMT) Conference*. Leuven, Belgium, pp. 161–168.
URL http://www.mt-archive.info/EAMT-2011-Potet.pdf
- [POT12] Potet, M., Esperança-Rodier, E., Besacier, L. and Blanchon, H. (2012). Collection of a Large Database of French-English SMT Output Corrections. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey, pp. 4043-4048.
URL www.lrec-conf.org/proceedings/lrec2012/pdf/506_Paper.pdf
- [QTLP] QT LaunchPad Project
URL http://www.qt21.eu/launchpad/
- [ROS12] Rosa, R., D. Mareček and O. Dusek (2012). DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the ACL Seventh Workshop on Statistical Machine Translation (WMT-2012)*. Montreal, Canada, pp. 362–368.
URL http://www.statmt.org/wmt12/pdf/WMT46.pdf
- [SIM07] Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical Phrase-based Post-editing. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).* Rochester, NY, pp. 508–515.
URL http://www.aclweb.org/anthology/N07-1064
- [SPE09] Specia, L., Turchi, M., Cancedda, N., Dymetman, M. and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT-2009)*. Barcelona, Spain, pp. 28-35.
URL http://www.mt-archive.info/EAMT-2009-Specia.pdf
- [SPE10] Specia, L., Cancedda, N. and Dymetman, M. (2010). A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC10).* Valletta, Malta, pp. 3375–3378.
URL http://clg.wlv.ac.uk/papers/Specia_LREC2010.pdf
- [SPE11a] Specia, L. (2011). Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation.* Leuven, Belgium, pp. 73-80
URL http://www.mt-archive.info/EAMT-2011-Specia.pdf

References

- [SPE11b] Specia, L., Hajlaoui, N. Hallett, C. and Aziz, W. (2011). Predicting Machine Translation Adequacy. *In Proceedings of The Thirteenth Machine Translation Summit (MTSummit-2011)*. Xiamen, China, pp 513--520.
  URL http://clg.wlv.ac.uk/papers/speciaetal.pdf
- [STY11] Stymne, S. (2011). Blast: A Tool for Error Analysis of Machine Translation Output. In *Proceedings of the ACL-HLT 2011 System Demonstrations*. Oregon, USA, pp. 56-61.
  URL http://www.aclweb.org/anthology/P11-4010
- [STY12] Stymne, S., and Ahrenberg, L. (2012). On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, pp. 1786-1790
  URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/717_Paper.pdf
- [TRA5] Translate5.
  URL http://www.translate5.net/
- [TUR13] Turchi, M., Negri, M. and Federico, M. (2013). Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT-2013)*. Sofia, Bulgaria, pp. 240-251.
  URL www.statmt.org/wmt13/pdf/WMT31.pdf

- [TUR13] Turchi, M., Negri, M. and Federico, M. (2013). Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT-2013)*. Sofia, Bulgaria, pp. 240-251.
  URL www.statmt.org/wmt13/pdf/WMT31.pdf

- [TUR14a] Turchi, M. and Negri, M. (2014). Automatic Annotation of Machine Translation Datasets with Binary Quality Judgements. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland.
  URL www.lrec-conf.org/proceedings/lrec2014/pdf/473_Paper.pdf
- [TUR14b] Turchi, M., Anastasopoulos, A., de Souza, J.G.C. and Negri, M. (2014). Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Baltimore, Maryland, pp. 710-720
  URL http://www.aclweb.org/anthology/P14-1067

References

- [VIL06] Vilar, D., Xiu, J., D'Haro, L. and Ney H. (2006) Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy, pp 697-702.
  URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf
- [WIS14] Wisniewski, G., Kübler, N. and Yvon, F. (2014). A Corpus of Machine Translation Errors Extracted from Translation Students Exercises. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland.
  URL www.lrec-conf.org/proceedings/lrec2014/summaries/1115.html
- [WIS13] Wisniewski, G., Singh, A.K., Segal, N. and Yvon, F. (2013). Design and analysis of a large corpus of post-edited translations: quality estimation, failure analysis and the variability of post-edition. In Proceedings of Machine Translation Summit (MT Summit 2013). Nice, France, pp. 117–124.
- [ZEM11] Zeman, D., Fishel, M., Bojar, O. and Berka, J. (2011). Addicter: What Is Wrong with My Translations? The Prague Bulletin of Mathematical Linguistics, 96, 2011, pp.79-88
  URL http://ufal.mff.cuni.cz/pbml/96/art-zeman-fishel-berka-bojar.pdf

Our annotation exercise

- annotation of *one* MT *English* output with respect to its *post-edition*
- *two* different settings:
  - annotation from scratch
  - revision of existing annotations
- MT output missing words must be annotated in the reference sentence

NOW GO TO: **http://mtequal.fbk.eu**

Typical error annotation issues

- disagreement as to whether something constitutes an error or not
- classification is ambiguous:
  - not easy to determine in which particular error category some error exactly belongs
  - there are several possible interpretations of the errors and different ideas about optimal solutions
- scope of span-level annotation: annotators agree on the error type but disagree on the precise span of the error (e.g. word order)
- one word can be assigned to more than one error category
- guidelines: often insufficient to guide annotators when faced with unfamiliar issues