LREC 2014

MTE - Workshop on Automatic and Manual Metrics for Operational Translation Evaluation

Relating Translation Quality Barriers to Source-Text Properties

Federico Gaspari, **Antonio Toral**,

Arle Lommel, Stephen Doherty, Josef van Genabith, Andy Way

QTLaunchPad EU project (grant agreement no. 296347)

{ fgaspari, atoral } @ computing.dcu.ie

Outline of the talk

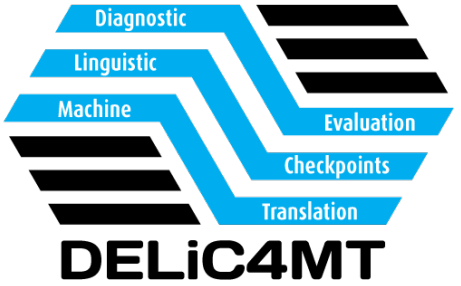
- Aim
- Translation quality barriers for MT
  - DELiC4MT diagnostic MT evaluation toolkit
- Study
  - Data, pre-processing and experimental set-up
  - Highlights of results
  - Analysis
- Conclusions
  - Summary and future work

Aim

- Identify source-side linguistic properties that pose MT quality barriers
  - For specific types of MT systems: SMT, RbMT, hybrid
  - For output of different quality levels: low, medium, high
  - Experiments for two language pairs: EN ↔ DE, EN ↔ ES

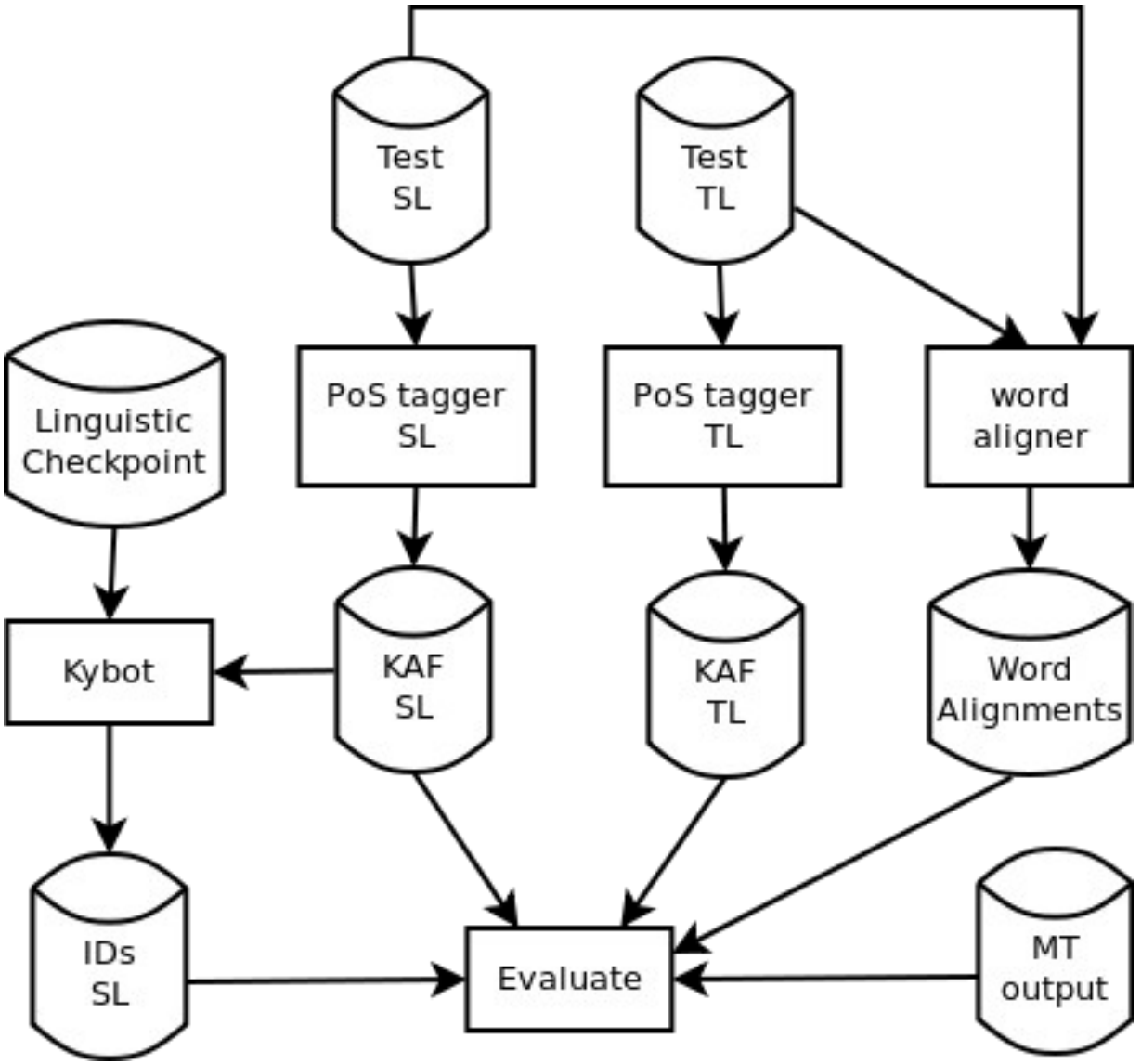
DELiC4MT

- **D**iagnostic **E**valuation using **Li**nguistic **C**heckpoints **for M**achine **T**ranslation
  - **www.computing.dcu.ie/~atoral/delic4mt**



- Linguistic checkpoints
  - Source-language phenomena that the user wants to investigate, e.g. PoS classes, lemmas, n-grams, literal words, any sequence
  - DELiC4MT score: ratio of correct MT output for SL checkpoints
- Strenghts of DELiC4MT
  - Open-source toolkit
  - Language-independent
  - Very flexible: any linguistic checkpoint, if features are supported

DELiC4MT architecture



DELiC4MT output: ES→EN, RbMT, verbs

- *Source:* Y aún así, <es> una estrella.
  Target ref: And yet, he <is> a star.
  MT output: And still like this, is a star.
  ngram matches: is (1/1)

- *Source:* Fue un regalo que me <hizo> él
  Target ref: It was a gift he <gave> me
  MT output: It was a gift that did me he
  ngram matches: - (0/1)

- *Source:* Anto tiene asma, <respira> con dificultad
  Target ref: Anto has asthma, <he> <has> difficulty breathing
  MT output: Anto has asthma, it breathes with difficulty
  ngram matches: has (1/3)

DELiC4MT and translation quality barriers

- So far DELiC4MT used for
  - Evaluation of overall MT output on user-defined checkpoints
- Novelty of this methodology
  - Application to the investigation of quality barriers in MT
- Relate translation quality barriers to source-text properties
  - DELiC4MT on subsets of MT output
  - Leading to comparative evaluations

Data sets

- News data
  - WMT 2013 data sets: "native" sentences with human reference

Translation Direction	Number of Sentences	MT Systems
EN→ES	500	SMT, RbMT, HMT
ES→EN	203	SMT, RbMT
EN→DE	500	SMT, RbMT, HMT
DE→EN	500	SMT, RbMT

  - *SMT:* phrase-based system by leading European research team
  - *RbMT:* well-established commercial system
  - *HMT (out of EN only):* well-established commercial system

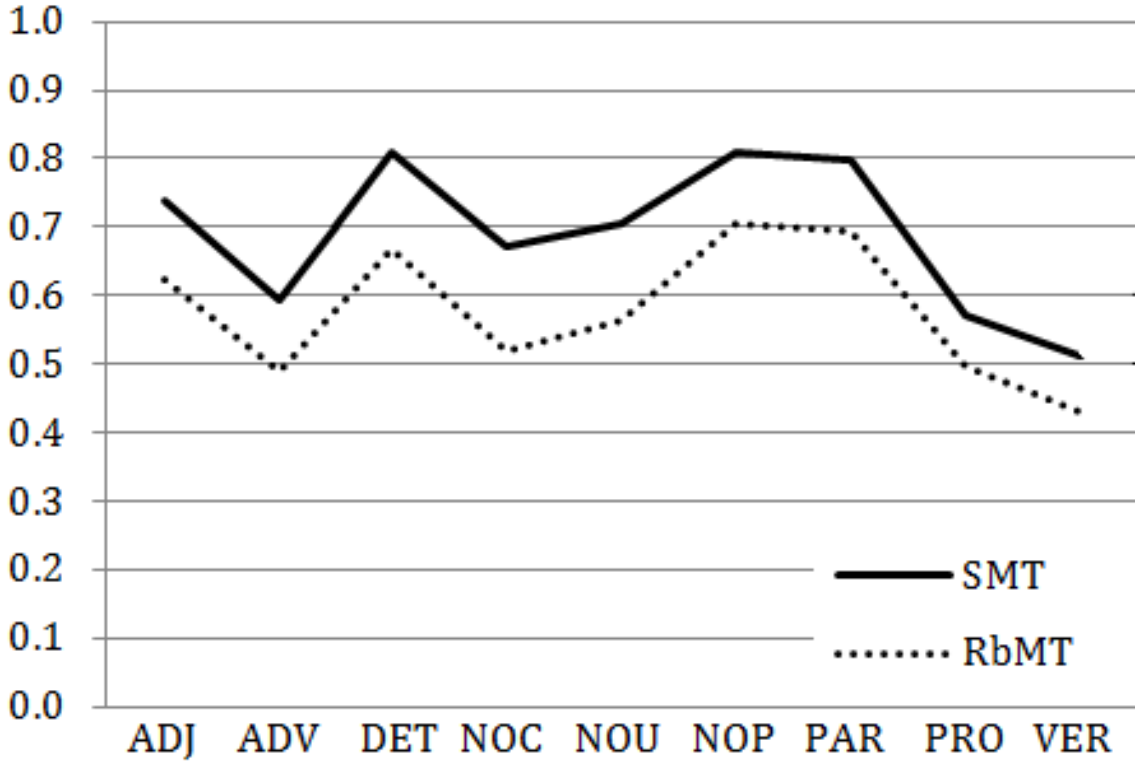
- SL linguistic checkpoints consist of 9 PoS classes, i.e.
  - ADJ, ADV, DET, NOC, NOP, NOU, PAR, PRO, VER

Pre-processing and set-up

- Two LSPs and a team of researchers evaluated MT output
  - *Good*: publishable, no post-editing required
  - *Near-miss*: less than 3 errors, easy to post-edit
  - *Poor*: 3 or more errors, requiring time-consuming post-editing
- DELiC4MT pre-processing
  - Source and target sides of the references were PoS-tagged
    - *Freeling* for EN and ES, TreeTagger for DE
    - Word alignment with GIZA++
- Same input evaluated when MT systems compared overall
  - But different data sets when evaluating specific quality ranking(s)

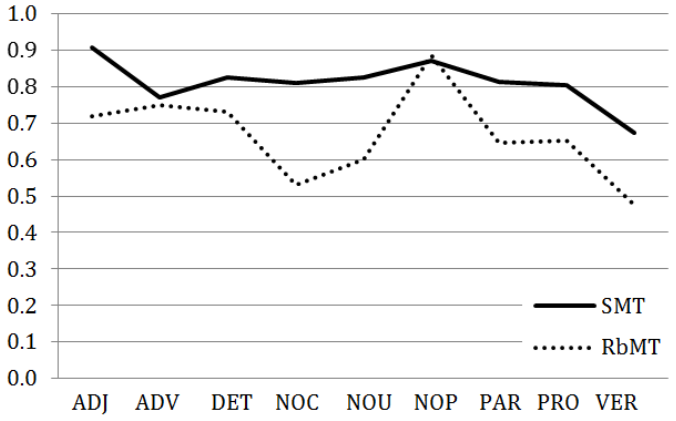
Evaluation: highlights of results

- ES → EN overall results (entire input)

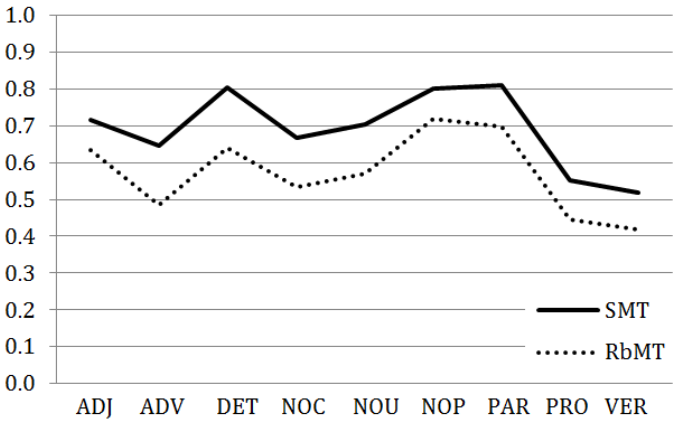


Evaluation: highlights of results

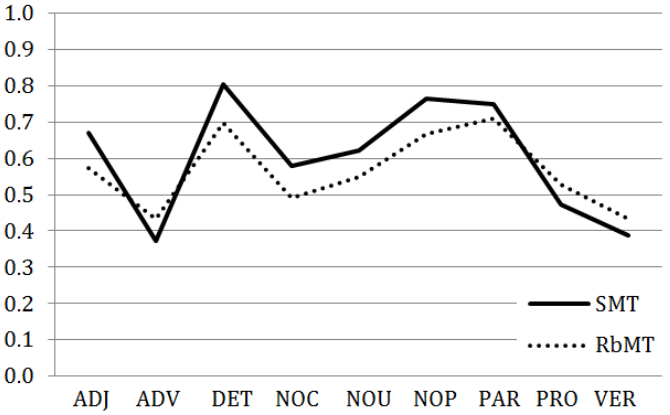
- ES → EN results for output broken down by quality ranking



Good-quality output



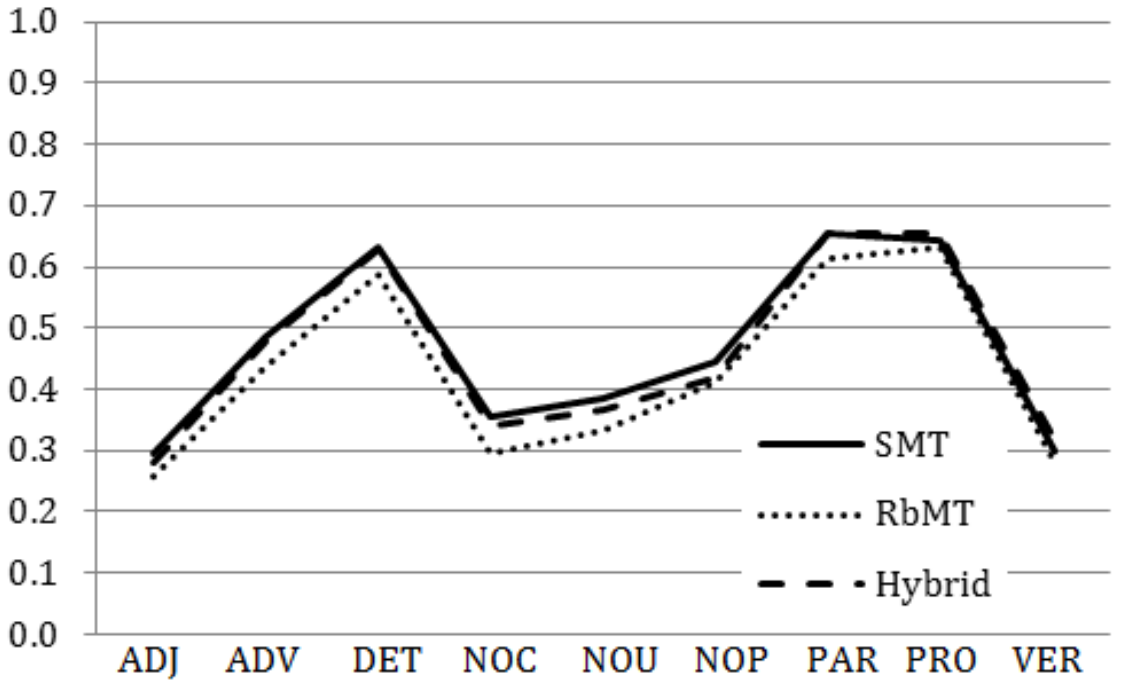
Near-misses



Poor-quality output

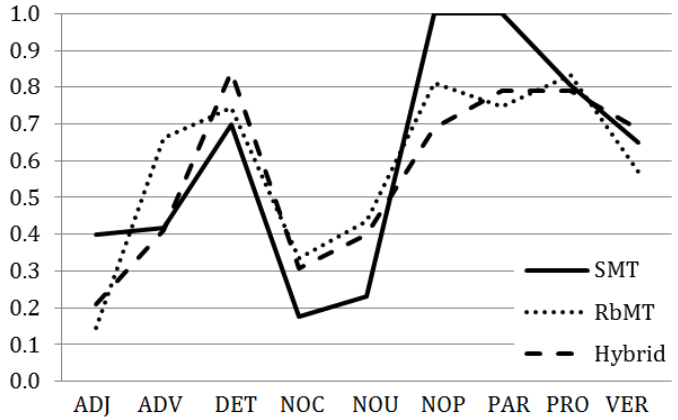
Evaluation: highlights of results

- EN → DE overall results (entire input)

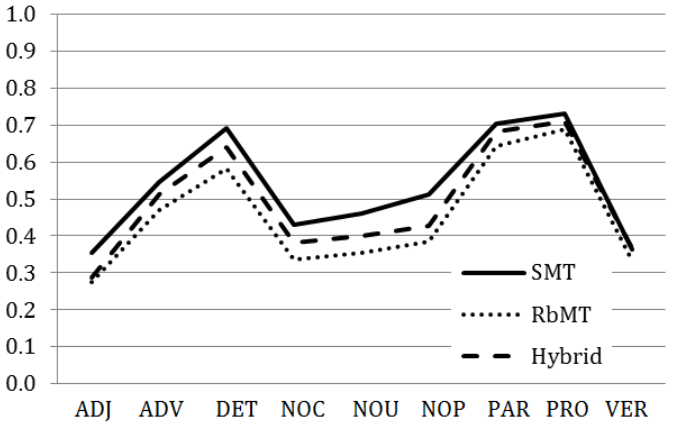


Evaluation: highlights of results

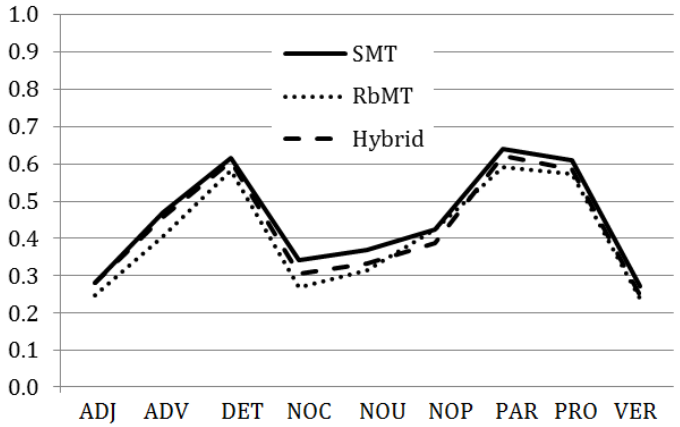
- EN → DE results for output broken down by quality ranking



Good-quality output



Near-misses



Poor-quality output

Analysis

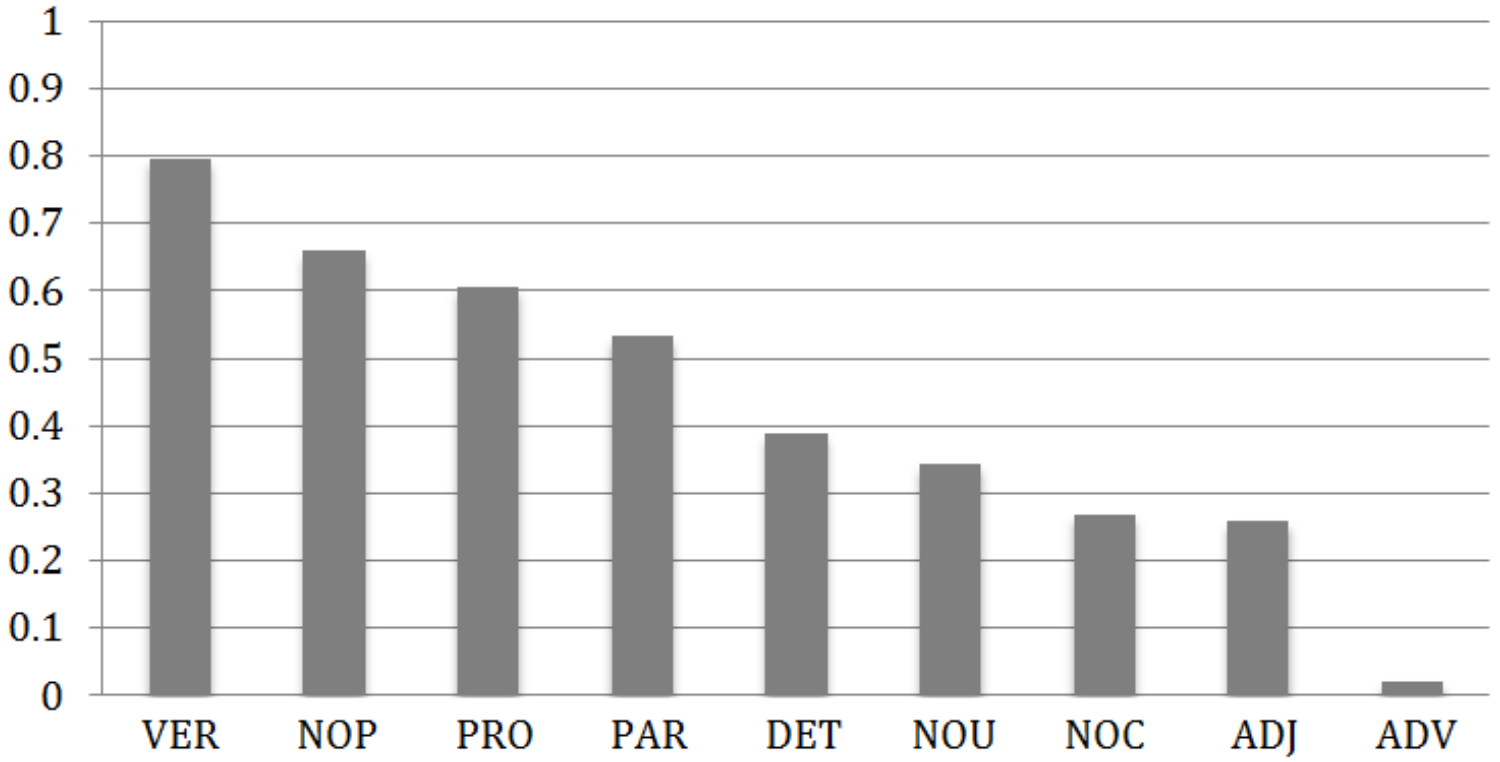
- Across all 4 translation directions, few "good" MT sentences
  - This leads to few linguistic checkpoints for the 9 PoS classes

		SMT		RbMT		HMT	
		ES>EN	EN>ES	ES>EN	EN>ES	ES>EN	EN>ES
GOOD	ADJ	53	59	27	44	-	71
	ADV	24	69	17	55	-	82
	DET	119	68	67	37	-	66
	NOC	193	197	100	97	-	183
	NOU	254	304	124	150	-	258
	NOP	61	107	24	53	-	75
	PAR	134	110	76	67	-	100
	PRO	41	56	29	48	-	70
	VER	177	193	157	102	-	198
NEAR-MISS	ADJ	196	492	196	496	-	448
	ADV	119	443	109	443	-	394
	DET	335	569	327	575	-	527
	NOC	639	1723	662	1772	-	1655
	NOU	853	2508	823	2499	-	2390
	NOP	214	785	161	727	-	735
	PAR	482	1156	459	1153	-	1073
	PRO	125	512	127	510	-	450
	VER	786	1449	687	1515	-	1373
POOR	ADJ	70	380	93	412	-	433
	ADV	51	302	68	327	-	352
	DET	132	340	188	373	-	393
	NOC	286	1202	354	1275	-	1308
	NOU	371	1649	528	1840	-	1852
	NOP	85	447	174	565	-	544
	PAR	163	729	240	793	-	845
	PRO	72	274	81	289	-	329
	VER	273	1016	388	1069	-	1117

		SMT		RbMT		HMT	
		DE>EN	EN>DE	DE>EN	EN>DE	DE>EN	EN>DE
GOOD	ADJ	173	5	45	20	-	38
	ADV	63	12	20	16	-	38
	DET	102	10	42	13	-	29
	NOC	356	40	152	44	-	140
	NOU	396	43	178	56	-	185
	NOP	39	3	26	12	-	45
	PAR	152	5	47	16	-	48
	PRO	87	16	35	18	-	30
	VER	179	40	89	42	-	75
NEAR-MISS	ADJ	591	180	587	360	-	419
	ADV	203	156	195	252	-	275
	DET	479	191	438	373	-	434
	NOC	2023	593	1655	1201	-	1312
	NOU	2294	905	1921	1878	-	1995
	NOP	270	312	264	677	-	683
	PAR	673	290	581	678	-	743
	PRO	298	167	278	317	-	336
	VER	680	493	640	896	-	1086
POOR	ADJ	536	708	673	512	-	436
	ADV	203	488	254	388	-	343
	DET	403	749	504	560	-	487
	NOC	1737	2551	2299	1930	-	1732
	NOU	1995	3771	2574	2766	-	2539
	NOP	258	1220	275	836	-	807
	PAR	581	1435	778	1034	-	939
	PRO	251	509	323	355	-	326
	VER	578	1801	708	1396	-	1173

Analysis

- Correlation between DELiC4MT scores and human evaluation
  - Pearson's *r* values for DELiC4MT scores and human ratings, broken down according to the 9 PoS-based linguistic checkpoints



Conclusions

- Applied DELiC4MT to the identification of source-side causes of MT quality barriers
  - 2 bidirectional language pairs: ES ↔ EN, DE ↔ EN
  - 3 types of MT systems: statistical, rule-based, hybrid
  - 3 quality levels of MT output: poor, near-miss, good
- Evaluation focused on 9 PoS-based linguistic checkpoints
  - Best quality predictors: VER, NOP and PRO; worst one is ADV
- Limitations
  - Few checkpoints detected for good MT output, data sparseness

Future work

- Analyse larger and different data sets with more linguistic checkpoints, sparseness
- Apply to new language pairs
- Explore connections with Multidimensional Quality Metric (MQM)
  - Potential of combining automatic diagnostic evaluation approaches with manual translation quality annotation
  - Relate specific MQM issue types to source-language properties

Relating Translation Quality Barriers to Source-Text Properties

Thank you for your attention!

(& we look forward to hands-on collaboration after lunch!)

Federico Gaspari, **Antonio Toral**,

Arle Lommel, Stephen Doherty, Josef van Genabith, Andy Way

QTLaunchPad EU project (grant agreement no. 296347)

{ fgaspari, atoral } @ computing.dcu.ie

References (1/2)

Balyan, R., Naskar, S.K., Toral, A. and Chatterjee, N. (2012). A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), COLING 2012*. Mumbai, India, December 2012, pp. 61--72.

Balyan, R., Naskar, S.K., Toral, A. and Chatterjee, N. (2013). A Diagnostic Evaluation Approach for English to Hindi MT Using Linguistic Checkpoints and Error Rates. In A. Gelbukh (Ed.), *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*. Samos, Greece. 2013. LNCS 7817. Berlin: Springer, pp. 285--296.

Burchardt, A., Gaspari, F., Lommel, A., Popović, M., and Toral, A. (2014). *Barriers for High-Quality Machine Translation*. QTLaunchPad Deliverable 1.3.1. Available from www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-1_3_1.pdf (accessed 10 February 2014).

Lommel, A. and Uszkoreit, H. (2013). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Paper presented at *Localization World*, 12-14 June 2013, London, United Kingdom.

Naskar, S.K., Toral, A., Gaspari, F. and Way, A. (2011). A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints. In *Proceedings of Machine Translation Summit XIII*. Xiamen, China, 19-23 September 2011, pp. 529--536.

Naskar, S.K., Toral, A., Gaspari, F. and Groves, D. (2013). Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation. In K. Sima'an, M.L. Forcada, D. Grasmick, H. Depraetere and A. Way (Eds.),*Proceedings of the XIV Machine Translation Summit*. Nice, France, 2-6 September 2013. Allschwil: The European Association for Machine Translation, pp. 135--142.

References (2/2)

Och, F.J., and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics,* volume 29, number 1, pp. 19--51.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference .* ELRA. Istanbul, Turkey. 21-27 May 2012, pp. 2473--2479.

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*. Philadelphia, July 2002, pp. 311--318.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. P*roceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland, pp. 47--50.

Toral, A., Naskar, S.K., Gaspari, F. and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics*, 98(1), pp. 121--131.

Toral, A., Naskar, S.K., Vreeke, J., Gaspari, F. and Groves, D. (2013). A Web Application for the Diagnostic Evaluation of Machine Translation over Specific Linguistic Phenomena. In C. Dyer and D. Higgins (Eds.), *Proceedings of the 2013 NAACL HLTConference - Demonstration Session.* Atlanta, GA, USA. 10-12 June 2013. Stroudsburg, PA: Association for Computational Linguistics, pp. 20--23.