Human Translation Evaluation

GF meets SMT Workshop
Gothenburg Nov 2, 2010

Maarit Koponen
University of Helsinki
maarit.koponen@helsinki.fi

Topics

1. Evaluation of human translation - Human evaluation of machine translation
2. Overview of manual evaluation methods
3. Example of an error analysis

Evaluation of Human and Machine translation

Human translation

- quantitative microtextual
  - Sical, ATA, MeLLANGE
- qualitative macrotextual
  - functional-pragmatic (House)
  - skopos-oriented (Nord)

Machine translation

- evaluation scales
- ranking
- error analysis
- reading comprehension
- post-editing

Evaluation scales

- fluency: How fluent is the candidate sentence?

This Green Paper seeks to bring all interested parties on the key issues that affect the formulation of future regulation.

5 flawless 4 good 3 non-native 2 disfluent 1 incomprehensible

- adequacy: How much of the information is present?

The purpose of this Green Paper is therefore to seek the views of all interested parties on the key issues that will shape the future Regulation.

This Green Paper seeks to bring all interested parties on the key issues that affect the formulation of future regulation.

5 all 4 most 3 much 2 little 1 none

(LDC 2005)

Evaluation scales (cont'd)

- utility: How effectively is the information conveyed?

Tällä vihreällä kirjalla pyritään saamaan kaikkien asianomaisten osapuolten näkemykset keskeisistä kysymyksistä, jotka vaikuttavat tulevan asetuksen muotoiluun.

This Green Paper seeks to bring all interested parties on the key issues that affect the formulation of future regulation.

4 complete 3 useful 2 marginal 1 poor

(TAUS 2006)

- clarity: How clear is the meaning of the sentence?

This Green Paper seeks to bring all interested parties on the key issues that affect the formulation of future regulation.

3 perfectly clear 2 1 0 not decipherable

(Miller and Vanni 2005)

Ranking

- pair-wise: Which is better?

This Green Paper seeks to bring all interested parties on the key issues that affect the formulation of future regulation.

With this Green Paper is designed to help views of interested parties of the key issues that affect future regulation wording.

(Vilar et al 2007)

- multiple: Rank from best to worst

With this Green Paper is designed to help views of interested parties of the key issues that affect future regulation wording.

This Green Paper seeks to bring all interested parties on the key issues that affect the formulation of future regulation.

Tällä green written pyritään indolent everybody asianomaisten side näkemykset keskeisistä kysymyksistä , which spectacular future asetuksen design.

(Callison-Burch et al 2007)

Post-editing

- Edit to publication quality

Kitchen furniture offer customers had asked the store in invoice design work. Client had received a decrease inappropriate, since the bid was not Charge told.

- Edit as necessary for understanding

Kitchen furniture offer customers had asked the store in invoice design work. Client had received a decrease inappropriate, since the bid was not Charge told.

- candidate

(Doyon et al 2008, Callison-Burch et al 2010)

Error analysis

- Subjective Sentence Error Rate SSER (Niessen et al 2000)

Kitchen furniture offer customers had asked the store in invoice design work.

- rate from 0 (nonsense) to 10 (perfect)
- divide into segments (information items)
- mark each information item as
- ok – missing – syntax – meaning – other

Reading comprehension

Kitchen furniture offer customers had asked the store in invoice design work. Client had received a decrease inappropriate, since the bid was not Charge told.

The customer

a) considered the invoice unnecessary

b) checked the price

c) did not want to use the plan

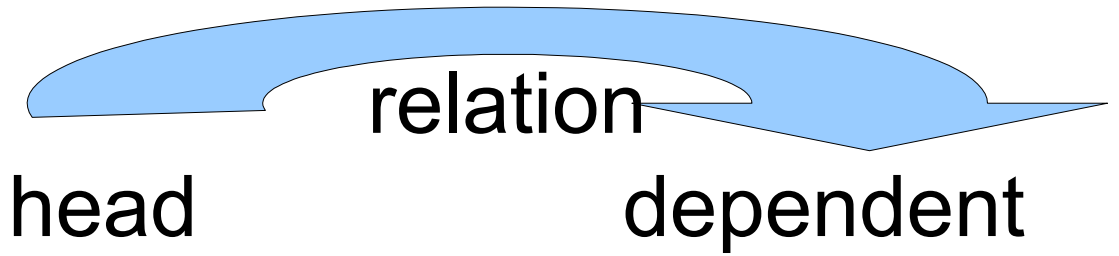
- percentage of correct answers between 55% … 76% (Tomita 1993, Fuji 1999)
- test subjects understand more than they expect (Fuji 1999)

Example of error analysis

- material:
  - European Commission Green Paper
  - Antivirus software installation guide
  - ~ 400 word passage analyzed
- statistical translation systems
  - statistical: Google, Bing
  - rule-based: Sunda, SDL Trados

Error analysis

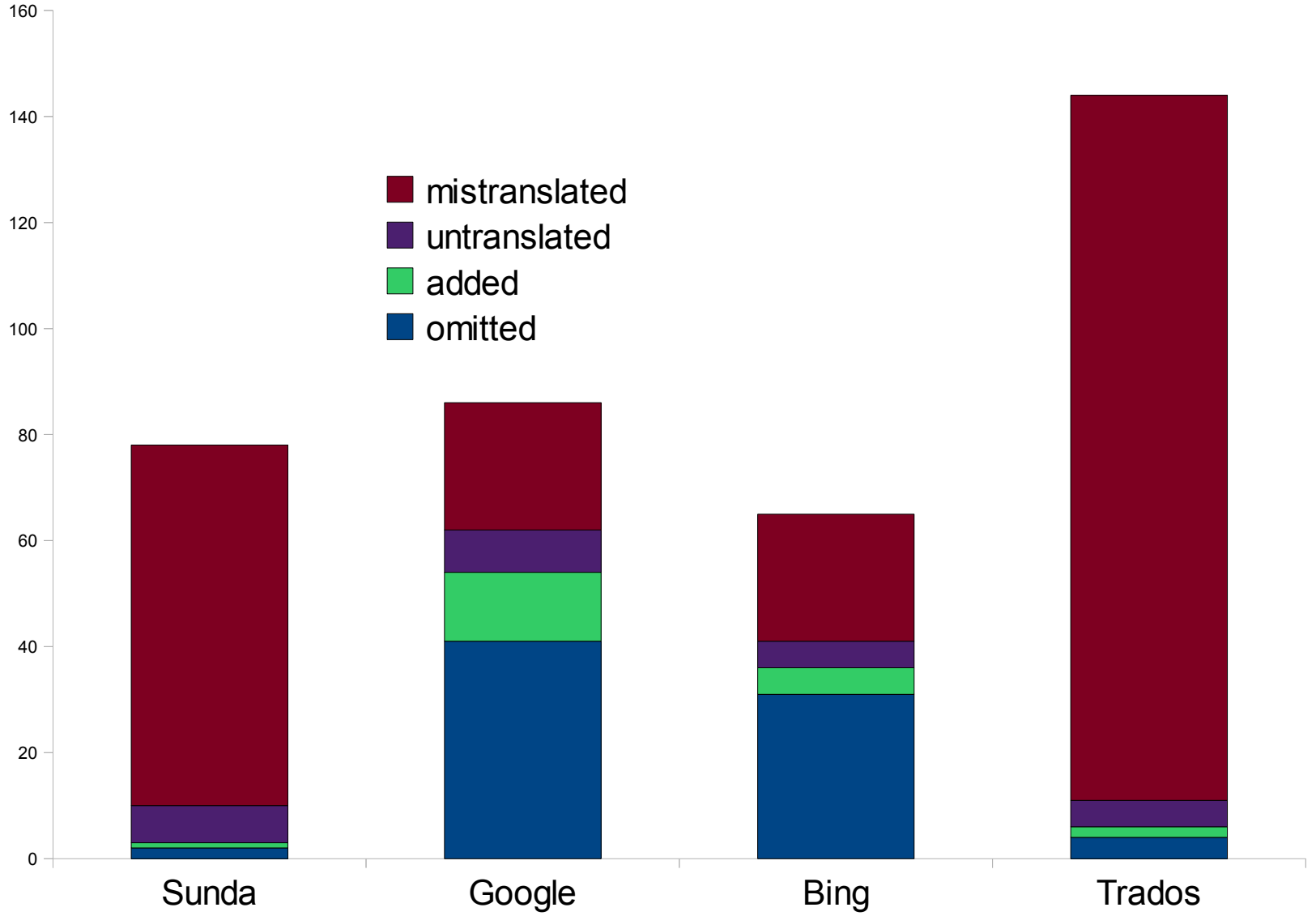
- language error vs. translation error
- definition of translation error:
  - semantic component not shared by ST and TT
  - concepts
  - relations



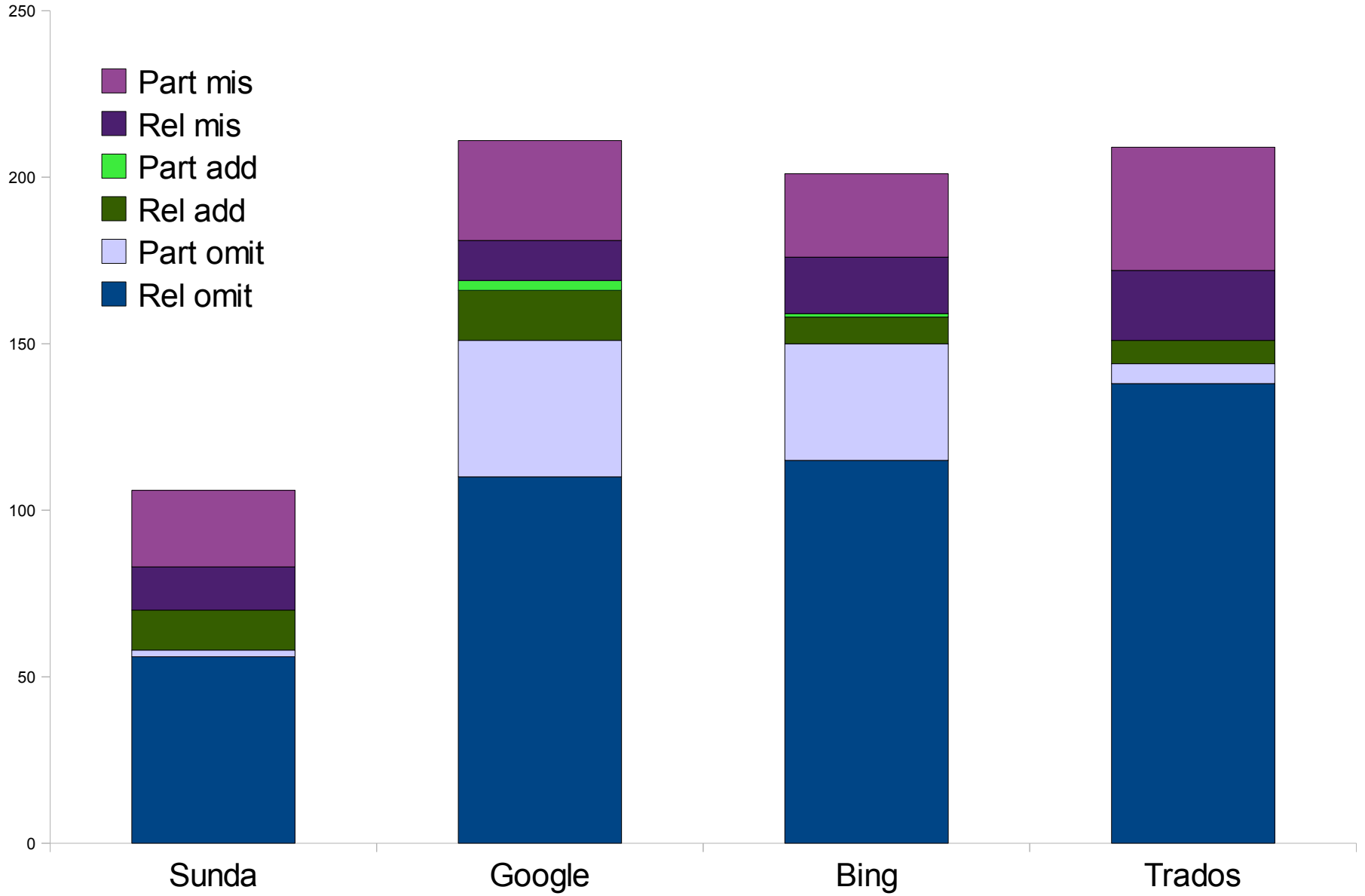
Error types

- concepts
  - omitted,
  - added
  - mistranslated
  - untranslated
- relations
  - omitted (relation/participant)
  - added (relation/participant)
  - mistaken (relation/participant)

Concept errors



Relation errors



Error rates vs. BLEU scores

	Concepts			**BLEU-1**		**Relations**			**BLEU-4**	
	Prec.	Rec.	F2	plain	tagged	Prec.	Rec.	F2	plain	tagged
Sunda EU	0.8686	0.8718	0.8712	0.3500	0.4901	0.8393	0.7550	0.7705	0.0950	0.1919
Sunda NAV	0.8475	0.8475	0.8475	0.3962	0.5317	0.8934	0.8289	0.8410	0.0787	0.2084
Sunda total	**0.8581**	**0.8597**	**0.8593**	**0.3731**	**0.5109**	**0.8664**	**0.7920**	**0.8058**	**0.0869**	**0.2002**
Google EU	0.8962	0.8535	0.8617	0.4985	0.6209	0.7487	0.5622	0.5917	0.1851	0.3127
Google NAV	0.8676	0.8369	0.8429	0.5642	0.6920	0.8385	0.6122	0.6471	0.2176	0.4184
Google total	**0.8819**	**0.8452**	**0.8523**	**0.5314**	**0.6565**	**0.7936**	**0.5872**	**0.6194**	**0.2013**	**0.3656**
Bing EU	0.9288	0.9084	0.9124	0.4753	0.5871	0.8383	0.5622	0.6019	0.1472	0.2575
Bing NAV	0.9132	0.8582	0.8686	0.5623	0.6841	0.8794	0.6654	0.6994	0.2422	0.4189
Bing total	**0.9210**	**0.8833**	**0.8905**	**0.5188**	**0.6356**	**0.8589**	**0.6138**	**0.6507**	**0.1947**	**0.3382**
Trados EU	0.6937	0.6886	0.6897	0.2959	0.4131	0.7865	0.5622	0.5963	0.0421	0.1148
Trados NAV	0.7908	0.7908	0.7908	0.4181	0.5637	0.8274	0.6198	0.6525	0.0732	0.2365
Trados total	**0.7423**	**0.7397**	**0.7402**	**0.3570**	**0.4884**	**0.8070**	**0.5910**	**0.6244**	**0.0577**	**0.1757**

Bibliography

- Human Translation

Bensoussan, Marsha, and Judith Rosenhouse. 1990. Evaluating student translations by discourse analysis. *Babel* 36. 65-84.

House, Juliane. 1981. *A model for translation quality assessment.* Tübingen: Narr.

House, Juliane. 2001. Translation quality assessment: Linguistic description versus social evaluation. *Meta* 46 (2). pp. 243-57.

Nord, Christiane. 2005. *Text analysis in translation: theory, methodology, and didactic application of a model for translation-oriented text analysis.* Amsterdam: Rodopi.

Waddington, Christopher. 2001. Different methods of evaluating student translations: The question of validity. *Meta* 46 (2). pp. 311-25.

Williams, Malcolm. 2001. The application of argumentation theory to translation quality assessment. *Meta* 46 (2). pp. 326-44.

- Error classifications

American Translators' Association. http://www.atanet.org/certification/aboutexams_presentation.php.

Schütz, Jörg. 1999. Deploying the SAE J2450 Translation Quality Metrics in Language Technology Evaluation Projects. *Aslib Conference Proceedings. Translationg and the Computer, 21.*

Secară, Alina. 2005. Translation evaluation: A State of the Art Survey. *Proceedings of the eCoLoRe/MeLLANGE Workshop.* 21-23 March 2005, Leeds, UK. pp. 39-44.

Bibliography

- Evaluation scales

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation.* June 23, 2007, Prague, Czech Republic. pp. 136-158.

Hamon, Olivier, Andrei Popescu-Belis, Khalid Choukri, Marianne Dabbadie, Anthony Hartley, Widad Mustafa El Hadi, Martin Rajman, and Ismail Timimi. 2006. CESTA: first conclusions of the Technolangue MT evaluation campaign. *LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings.* 22-28 May 2006, Genoa, Italy. pp. 179-184.

LDC. 2005. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations Revision 1.5, January 25, 2005*. Linguistic Data Consortium. http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf

Miller, Keith J., and Michelle Vanni. 2005. Inter-rater agreement measures, and the refinement of metrics in the PLATO MT evaluation paradigm. *Conference Proceedings: the tenth Machine Translation Summit.* September 13-15, 2005, Phuket, Thailand. pp. 125-132.

TAUS. 2006. *Quality evaluation and TA*. Translation Automation User Society. http://www.translationautomation.com/best-practices/quality-evaluation-and-ta.html

Bibliography

- Post-editing, ranking

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR.* July 2010. Uppsala, Sweden. pp. 17--53

Doyon, Jennifer, Christine Doran, C. Donald Means, and Domenique Parr. 2008. Automated Machine Translation Improvement Through Post-Editing Techniques: Analyst and Translator Experiments. *AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas.* 21-25 October 2008, Waikiki, Hawai'i. pp. 346-353.

Glenn, Meghan Lammie, Stephanie Strassel, Lauren Friedman, and Haejoong Lee. 2008. Management of large annotation projects involving multiple human judges: a case study of GALE machine translation post-editing. *LREC 2008: 6th Language Resources and Evaluation Conference.* 26-30 May 2008, Marrakech, Morocco.

Vilar, David, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation.* June 23, 2007, Prague, Czech Republic. pp. 96-103.

Bibliography

- Error analysis

Moré López, Joaquim, and Salvador Climent Roca. 2008. A machine translationness typology for MT evaluations. *Proceedings of the twelfth conference of the European Association for Machine Translation.* September 22-23, 2008, Hamburg, Germany. pp. 130-139.

Niessen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: fast evaluation for MT research. *LREC-2000: Second International Conference on Language Resources and Evaluation. Proceedings.* 31 May – 2 June 2000, Athens, Greece. pp. 39-45.

Popovic, Maja, and Hermann Ney. 2009. Syntax-Oriented Evaluation Measures for Machine Translation Output. *Proceedings of the Fourth Workshop on Statistical machine translation.* 30 March – 31 March 2009, Athens, Greece. pp. 29-32.

Popovic, Maja, Hermann Ney, Adrià de Gispert, José B. Mariño, Deepa Gupta, Marcello Federico, Patrik Lambert, and Rafael Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. *StatMT '06: Proceedings of the Workshop on Statistical Machine Translation.* 8-9 June, 2006, New York City, New York. pp. 1-6.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. *LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings.* 22-28 May 2006, Genoa, Italy. pp. 697-702.

Bibliography

- Reading comprehension

Fuji, Masaru. 1999. Evaluation experiment for reading comprehension of machine translation outputs. *Proceedings of MT Summit VII "MT in the Great Translation Era".* 13-17 September 1999, Singapore. pp. 285-289.

Jones, Douglas, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. *Proceedings of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* March 18-23, 2005, Philadelphia, PA, USA. pp. 1009-1012.

Jones, Douglas, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen, Edward Gibson, and Michael Emonts. 2007. ILR-based MT comprehension test with multi-level questions. *NAACL '07: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers on XX.* 22-27 April 2007, Rochester, New York. pp. 77-80.

Tomita, Masaru, Masako Shirai, Junya Tsutumi, Miki Matsumura, and Yuki Yoshikawa. 1993. Evaluation of MT systems by TOEFL. *TMI-93: The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation. Proceedings.* July 14-16, 1993, Kyoto, Japan. pp. 252-265.

Voss, Clare R., and Calandra R. Tate. 2006. Task-based evaluation of machine translation (MT) engines. Measuring how well people extract who, when, where-type elements in MT output. *EAMT-2006: 11th Annual Conference of the European Association for Machine Translation. Proceedings.* June 19-20, 2006, Oslo, Norway. pp. 203-212.