Automatic error analysis of machine translations

German
Research Center
for Artificial
Intelligence

Maja Popović

Errare 2013
(Paris, France)

21 November 2013

Motivation

- standard automatic evaluation metrics (BLEU, TER, METEOR) do not provide answers on questions such as:
  - what is a particular strength/weakness of the system?
  - what does a particular modification exactly improve?
  - does a worse-ranked system outperform a better-ranked one in any aspect?

→ human error analysis and classification have become widely used in recent years for these purposes

– human evaluation is resource-intensive and time-consuming

⇒ automatic error analysis

Automatic method for error classification

Two main goals of our automatic method:

- distribution of errors over the error classes within an output
- distribution of errors over translation outputs within a class

Five error classes based on the human error analysis scheme (Vilar+ 2006):

- inflectional errors
- reordering errors
- missing words
- extra words
- incorrect lexical choice

using WER and PER decomposition method

(Popović & Ney 2011, "Towards Automatic Error Analysis of Machine Translation Output", Computational Linguistics)

Definition of error classes

After identifying actual words contributing to the:

- Levenshtein distance WER
- reference position-independent error rate RPER
- hypothesis position-independent error rate HPER

the error classes are defined:

- inflectional error:
  full form is an RPER or HPER error, base form is correct
- reordering error:
  a WER error which is neither RPER nor HPER error
- missing word:
  a WER deletion which is also an RPER error
- extra word:
  a WER insertion which is also an HPER error
- lexical error:
  an error which is neither inflectional nor missing/extra word

Examples

- inflectional error
- word order error
- missing word/extra word
- lexical error

reference: the total amount designated for assistance to the system is to be divided into two parts .
MT output: the for the system to help certain total amount will be divided into two parts .

reference: there are price and qualitative categories here as well .
MT output: here too there is price and quality categories .

reference: one of the most beautiful national anthems .
MT output: one of the finest anthem .

`http://www.dfki.de/~mapo02/hjerson/`

(Popović 2011, ”Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output”)

Correlation with human error classification results

Spearman's and Pearson's correlation coefficients ρ and r

- across error classes in one translation output

output	ρ_{output}/r_{output}
EsEn1	0.90/0.92
EsEn2	0.90/0.91
EsEn3	0.90/0.98
EsEn4	0.90/0.98
EsEn5	0.90/0.99
EsEn6	0.90/0.99
DeEn1	0.70/0.72
DeEn2	0.70/0.74
DeEn3	0.90/0.91

output	ρ_{output}/r_{output}
ArEn1	0.90/0.96
ArEn2	1.00/0.99
CnEn	1.00/0.93
EsEn1'	0.80/0.94
EsEn2'	0.80/**0.55**
EsEn3'	0.80/0.98
EnEs1	0.95/0.75
EnEs2	**0.60**/**0.57**
EnEs3	1.00/**0.54**

- across different translation outputs

ρ_{class}/r_{class}	inflection	order	missing	extra	lexical
EnEs1-6	0.94/0.99	0.83/0.85	0.87/0.99	**-0.19/-0.34**	0.99/0.99
DeEn1-3	1.00/0.90	1.00/0.99	0.60/0.90	0.50/0.62	1.00/0.96

Recall and precision (DeEn1)

reference	inflection	order	missing	lexical	x
inflection	**92.3**/37.5	1.6/3.1	2.0/12.5	1.6/9.4	1.1/37.5
order	/	**61.3**/15.3	5.9/4.8	2.6/2.0	17.3/**77.8**
missing	/	6.5/2.1	45.8/48.4	16.6/16.7	5.7/32.8
lexical	7.7/0.2	11.3/1.4	42.9/17.5	**78.2**/30.3	22.6/**50.6**
x	/	19.4/1.9	3.4/1.1	1.0/0.3	**53.4/96.6**

MT output	inflection	order	extra	lexical	x
inflection	**92.3**/37.5	5.4/12.5	/	2.6/12.5	1.1/37.5
order	/	**51.4**/15.3	14.8/3.2	4.5/2.8	17.8/**78.6**
extra	/	1.4/3.2	16.7/29.0	3.2/16.1	1.5/**51.6**
lexical	7.7/0.2	24.3/4.0	**57.4**/6.9	**85.8**/29.6	24.4/**59.3**
x	/	17.6/2.1	11.1/1.0	3.9/1.0	**55.3/96.0**

- lower precisions ⇔ lower recall of correct words
  - especially for reordering and lexical errors
- large confusion “missing/extra words ⇒ lexical errors”
- a number of frequent extra words is
  - not detected
  - tagged as reordering errors

Comparison with human classification – summary

- a systematic method for automatic error classification
  - high correlations with human classification results
  - high recall values*

  $\Rightarrow$ can replace or facilitate human error analysis

*except extra words

$\rightarrow$ not particularly stable and reliable at this stage

recent experiment:

- calculating WER on base forms improves recall and precision

Applications: analysing a MT system

- analysing errors of a machine translation system (statistical phrase-based system)

N_{errors}	infl	order	missing	extra	lexical
DeEn1	32	235	199	40	521

  - predominant type of errors is lexical
  - followed by reordering errors and missing words
  - * can I do something about it?

    → attack on the reordering errors:

    - apply POS-based German verb pre-reordering for the original system (DeEn2)
    - using a hierarchical phrase-based system (DeEn3)

Applications: comparing MT systems

- comparing different translation systems (or variations)

  $\rightarrow$ introducing error rates $N_{errors}/N_{words}(\%)$

* what have I done?

error rates (%)	infl	order	missing	extra	lexical
DeEn1	2.1	14.8	12.5	2.5	32.8
DeEn2	2.8	**13.3**	12.6	3.5	**31.2**
DeEn3	2.9	17.2	**9.6**	6.2	**32.0**

  - DeEn2: verb reordering reduced reordering (explicitly) and lexical (implicitly) errors
  - DeEn3: hierarchical system[1] reduced missing words and lexical errors

[1] a very old version of hierarchical system

Applications: involving human translators

Analysis of post-edit operations performed by human translators

- post-edited translation output = reference translation

original:	all functions can be stopped with the red control knob , activate or deactivate .
post-edited:	all functions can be stopped , activated or deactivated , with the red control button .
original:	installation and electrical connection only by technical personnel and in accordance with valid regulations would drive through leave !
post-edited:	installation and electrical connection to be carried out only by qualified personnel and in accordance with valid regulations !
original:	danger of property damages !
post-edited:	danger of material damage !

Analysis of post-edits: selection

Selecting one of the several offered machine translation outputs for post-editing

* are there some (less) preferred edit operations?

	edit rates (%) selected	rest	relative difference (%)
form	2.9	4.5	36.2
order	5.3	7.8	31.9
missing	3.6	6.7	**45.8**
extra	6.0	9.0	34.2
lexical	21.2	33.0	35.8

* selection vs. ranking

total edit rate (%)	selected rank≠1	not selected rank 1
de-en	30.8	38.9
de-es*	**35.9**	33.9
de-fr	57.8	67.8
en-de*	**44.9**	37.6
es-de	32.8	51.7
fr-de	42.4	44.5

Analysis of post-edits: translation quality

Translation quality levels and post-editing operations (QTLAUNCHPAD[2] project)

* what are the differences between types of edit operations for different quality levels?

Following quality levels were assigned to the analysed translation outputs by human annotators:

- acceptable (ok)
- almost acceptable, easy to post-edit (edit+)
- possible to edit (edit)
- still possible to edit, better than from scratch (edit-)
- very low quality, better translate from scratch than try to post-edit (bad)

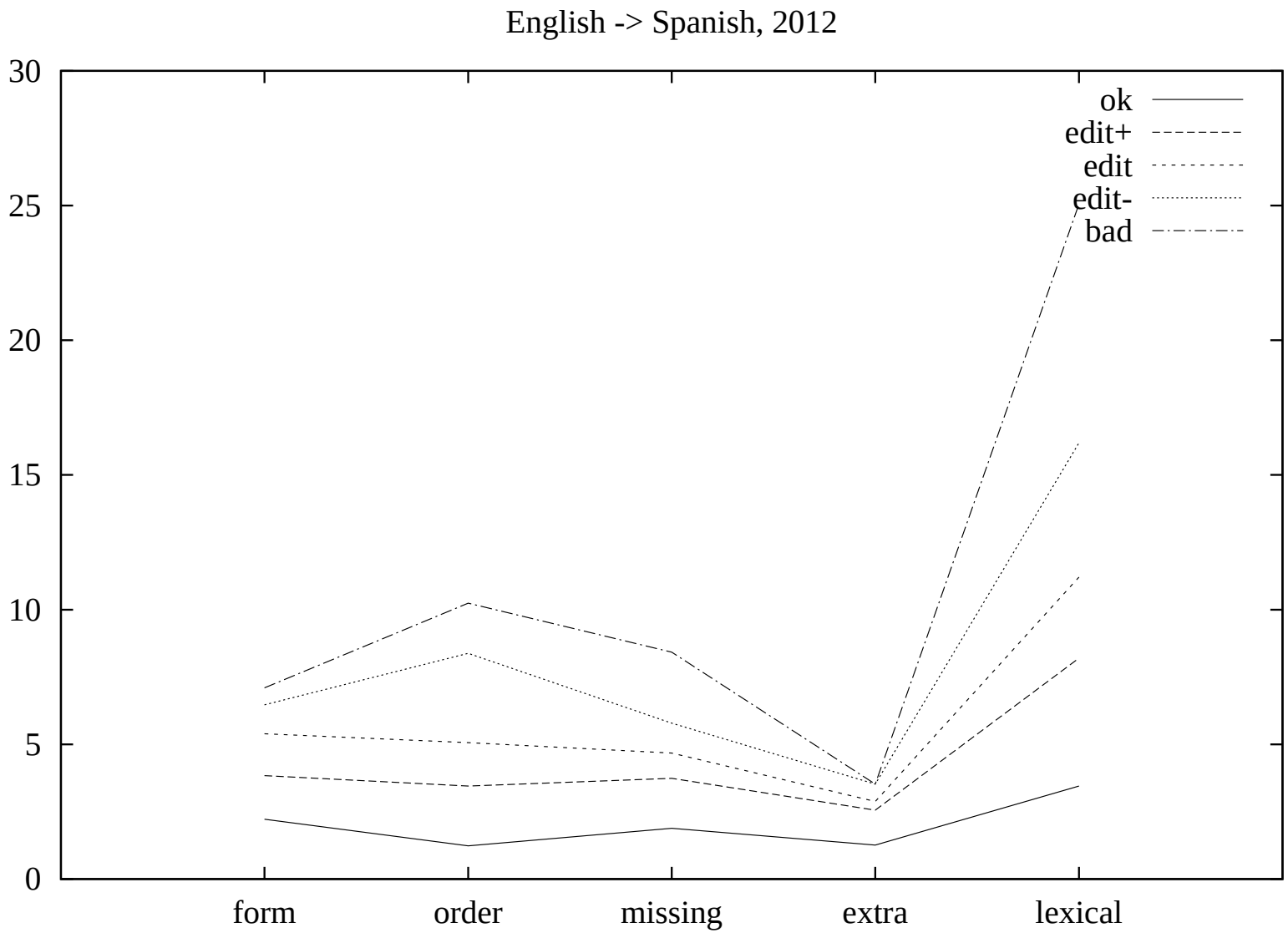
[2] http://qt21.eu/launchpad

WMT news data translated by statistical systems

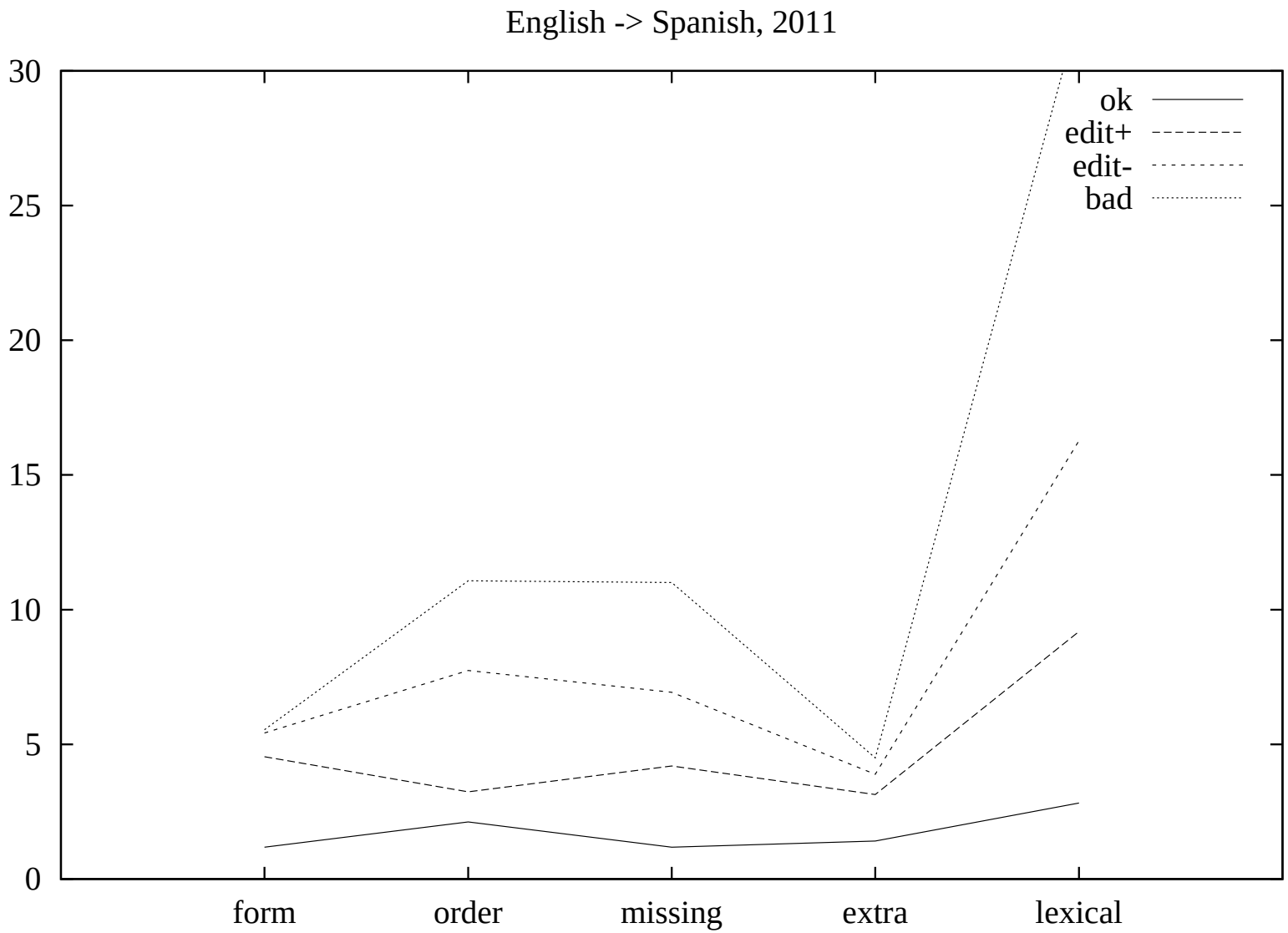
English→Spanish and French→English

number of sentences	total	ok	edit+	edit	edit-	bad
en-es 2012	2254	200	548	856	576	74
en-es 2011	1000	31	399	0	550	20
fr-en 2011	2525	323	1559	0	544	99

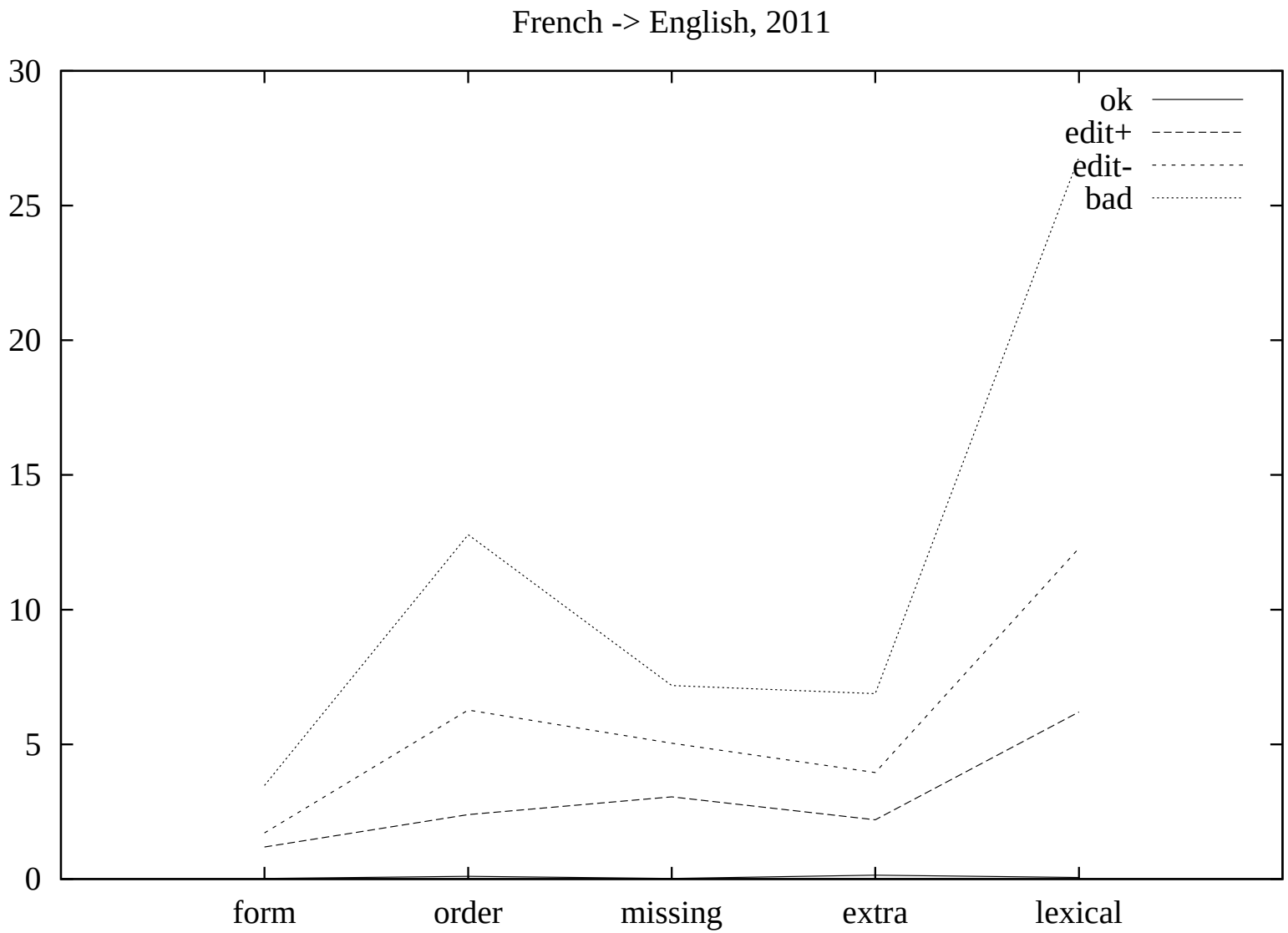
Distribution of edit rates (%) for each quality level



Distribution of edit rates (%) for each quality level



Distribution of edit rates (%) for each quality level



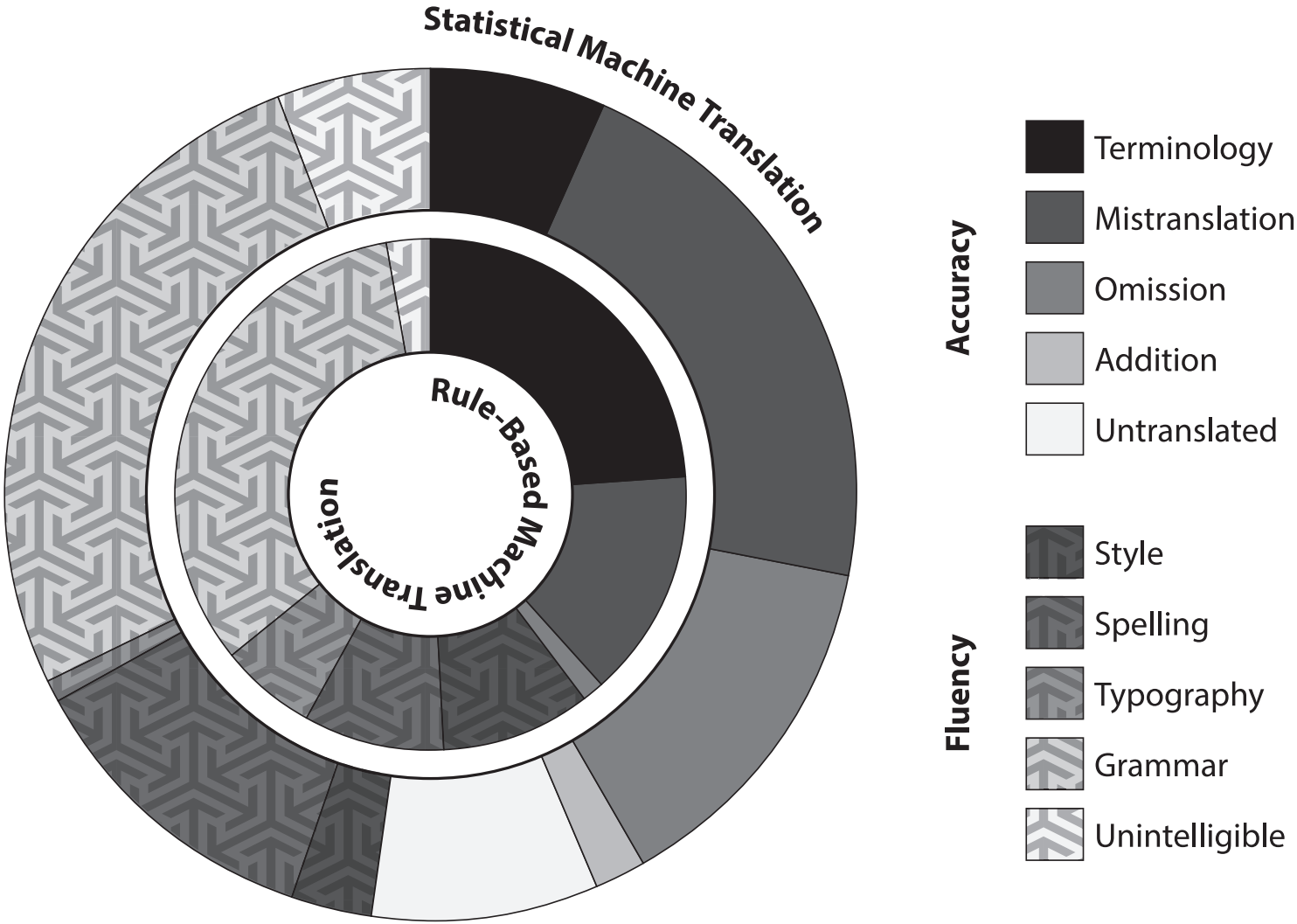
Multidimensional Quality Metric (MQM)

- Accuracy
  - Terminology ⇔ lexical choice
  - Mistranslation ⇔ lexical choice
  - Omission ⇔ missing word
  - Addition ⇔ extra word
  - Untranslated ⇔ lexical choice
- Fluency
  - Style
  - Spelling ∼ lexical choice
    - Capitalization ∼ lexical choice, inflection
  - Typography ∼ lexical choice
    - Punctuation ⇔ lexical choice
  - Grammar
    - Morphology (word form)
    - Part of speech ≈ lexical choice, inflection
    - Agreement ⇔ inflection
    - Word order ⇔ word order
    - Function words ∼ any class
  - Unintelligible

Multidimensional Quality Metric (MQM)

Accuracy		
Terminology		⇔ `lexical choice`
Mistranslation		⇔ `lexical choice`
Omission		⇔ `missing word`
Addition		⇔ `extra word`
Untranslated		⇔ `lexical choice`
Fluency		
Style		
Spelling		∼ `lexical choice`
	Capitalization	∼ `lexical choice, inflection`
Typography		∼ `lexical choice`
	Punctuation	∼ `lexical, missing/extra`
Grammar		
	Morphology	
	Part of speech	∼ `lexical choice, inflection`
	Agreement	⇔ `inflection`
	Word order	⇔ `word order`
	Function words	∼ `any class`
Unintelligible		

MQM results for a German→English text



Error distribution in QTLaunchPad pilot corpus

Automatic results for the same text

post-edited translation as reference

edit rates (%)	SMT	RBMT
inflection	1.4	0.9
word order	1.2	0.9
missing	4.4	1.4
extra	1.2	1.9
lexical	5.0	5.7

Some tendencies can be spotted immediately:

- much more omissions in SMT outputs
- lexical choice better handled by SMT systems
- grammar slightly better handled by RBMT systems

Summary

- results for different quality levels stable for all data sets:
  - no crossings in distributions of edit types
  - lexical and word order errors increase when quality decreases
- automatic analysis of edit types correlate with (rough) MQM results
- a number of possible directions for future work

* this work has been partly supported by the QTLAUNCHPAD (EU FP7 CSA No. 296347) project
* many thanks to Hans Uszkoreit and Arle Lommel