Quality Estimation for Machine Translation: different users, different needs

Lucia Specia

L.Specia@wlv.ac.uk

http://pers-www.wlv.ac.uk/~in1316/

JEC Workshop

October 14th 2011

UNIVERSITY OF WOLVERHAMPTON

Quality Estimation for Machine Translation: different translators, same needs

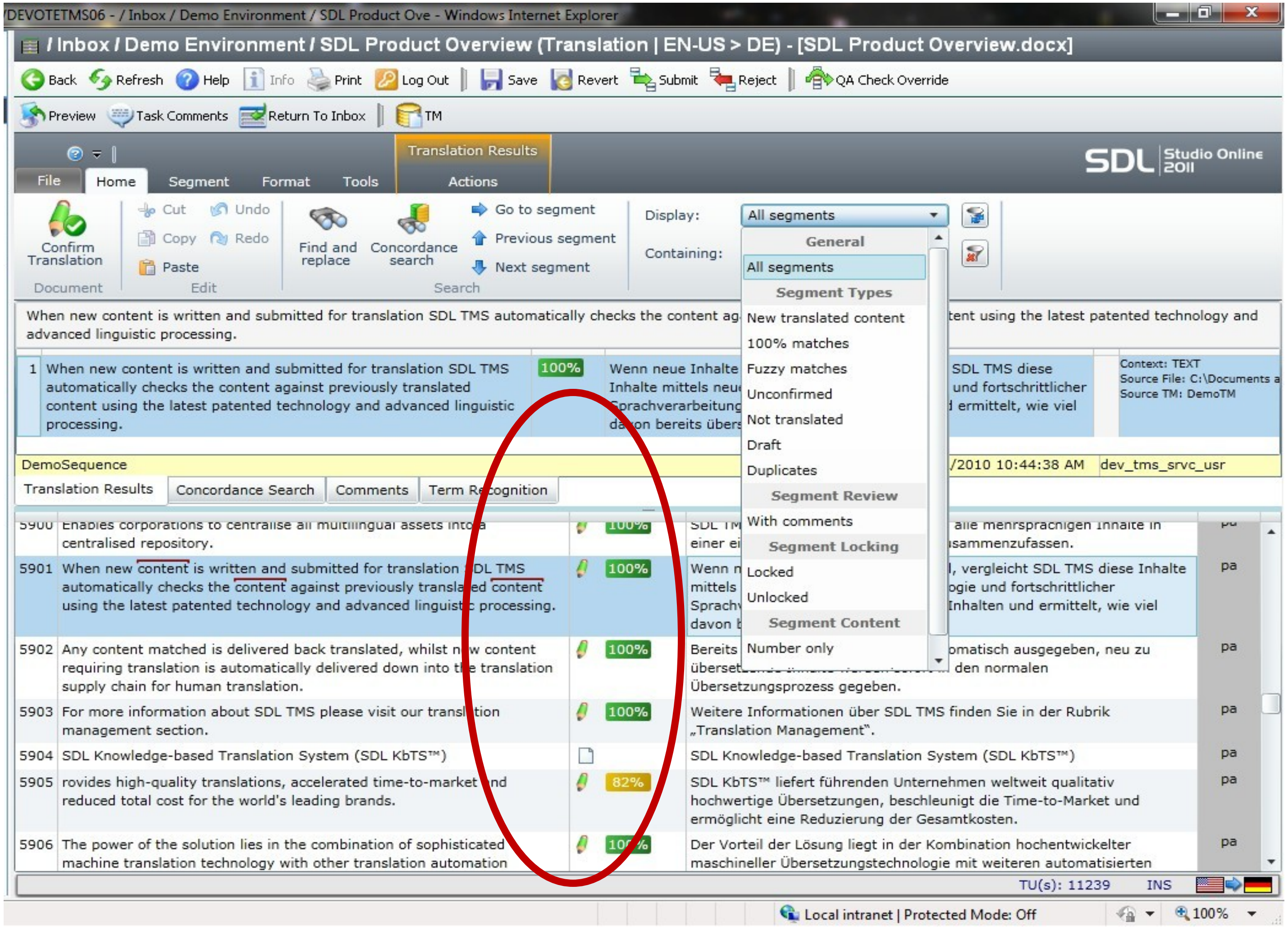
Lucia Specia
L.Specia@wlv.ac.uk
http://pers-www.wlv.ac.uk/~in1316/

JEC Workshop
October 14th 2011

UNIVERSITY OF WOLVERHAMPTON

Why are you not (yet) using MT?

- ❑ Why do you use translation memories?
- ❑ Perfect translations?



When new content is written and submitted for translation SDL TMS automatically checks the content ag … tent using the latest patented technology and advanced linguistic processing.

1	When new content is written and submitted for translation SDL TMS automatically checks the content against previously translated content using the latest patented technology and advanced linguistic processing.	100%	Wenn neue Inhalte … SDL TMS diese Inhalte mittels neu … und fortschrittlicher Sprachverarbeitung … ermittelt, wie viel davon bereits übers	Context: TEXT Source File: C:\Documents a Source TM: DemoTM

DemoSequence … /2010 10:44:38 AM dev_tms_srvc_usr

5900	Enables corporations to centralise all multilingual assets into a centralised repository.	100%	SDL TM … alle mehrsprachigen Inhalte in einer ei … usammenzufassen.	pa
5901	When new content is written and submitted for translation SDL TMS automatically checks the content against previously translated content using the latest patented technology and advanced linguistic processing.	100%	Wenn n … l, vergleicht SDL TMS diese Inhalte mittels … ogie und fortschrittlicher Sprachv … Inhalten und ermittelt, wie viel davon b	pa
5902	Any content matched is delivered back translated, whilst new content requiring translation is automatically delivered down into the translation supply chain for human translation.	100%	Bereits … omatisch ausgegeben, neu zu überset … den normalen Übersetzungsprozess gegeben.	pa
5903	For more information about SDL TMS please visit our translation management section.	100%	Weitere Informationen über SDL TMS finden Sie in der Rubrik „Translation Management".	pa
5904	SDL Knowledge-based Translation System (SDL KbTS™)		SDL Knowledge-based Translation System (SDL KbTS™)	pa
5905	rovides high-quality translations, accelerated time-to-market and reduced total cost for the world's leading brands.	82%	SDL KbTS™ liefert führenden Unternehmen weltweit qualitativ hochwertige Übersetzungen, beschleunigt die Time-to-Market und ermöglicht eine Reduzierung der Gesamtkosten.	pa
5906	The power of the solution lies in the combination of sophisticated machine translation technology with other translation automation	100%	Der Vorteil der Lösung liegt in der Kombination hochentwickelter maschineller Übersetzungstechnologie mit weiteren automatisierten	pa

Outline

- ❑ **Quality Estimation (QE)** for Machine Translation (MT)

- ❑ **Applications**

- ❑ General **approach**

- ❑ What **aspect of quality** we want to estimate and **how to represent it**

- ❑ How we **assess** quality estimation systems

QE for MT

- ❑ **Goal**: given the output of an MT system for a given input, provide an estimate of its **quality**

- ❑ **Motivations**: assessing the quality of translations is
  - ❑ **Time consuming, tedious, not worth it**

Une interdiction gouvernementale sur la non-UE conjoints étrangers de moins de 21 à venir au Royaume-Uni, qui a été introduit par le Labour en 2008 et vise un partenaire étranger de l'extérieur de l'UE ne pouvait pas se joindre à leurs partenaires au Royaume-Uni si elles étaient moins de 21 ans, est illégale, disent les juges haut.

  - ❑ **Not always possible**

个非欧盟国家的外国配偶来英国，□是在 2008 年由□工□推出，□意味着

QE for MT

- Main applications:

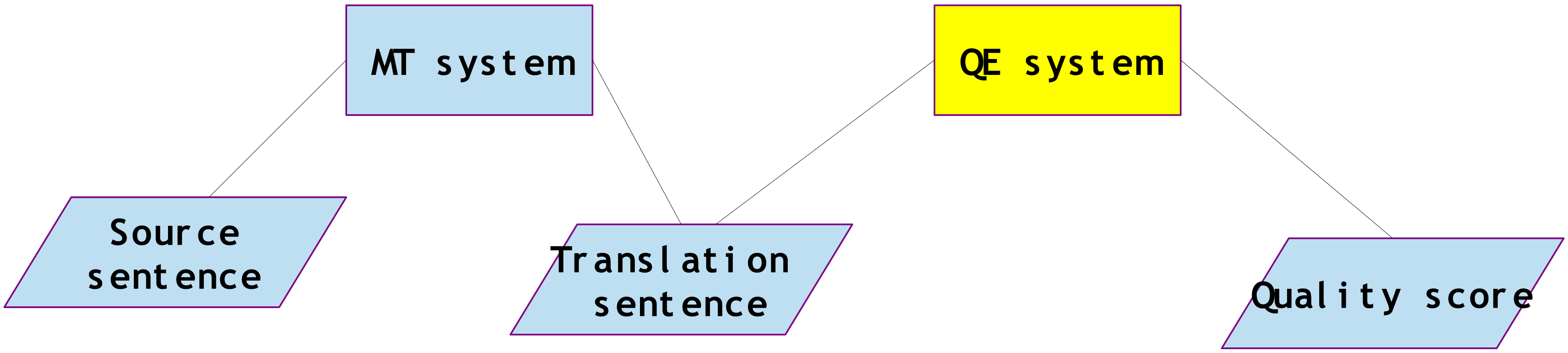
Is it worth providing this translation to a professional translator for post-editing?

Should this translation be highlighted as "not reliable" to a reader?

ven multiple translation options for a given input can we select the best one?

s this sentence good enough for publishing as is?

QE for MT



- **Different from MT evaluation (BLEU, NIST, etc):**
  - MT system in use, translating **unseen text**
  - **Translation unit: sentence** → not about **average** quality
  - **Independent** from MT system (post-MT)

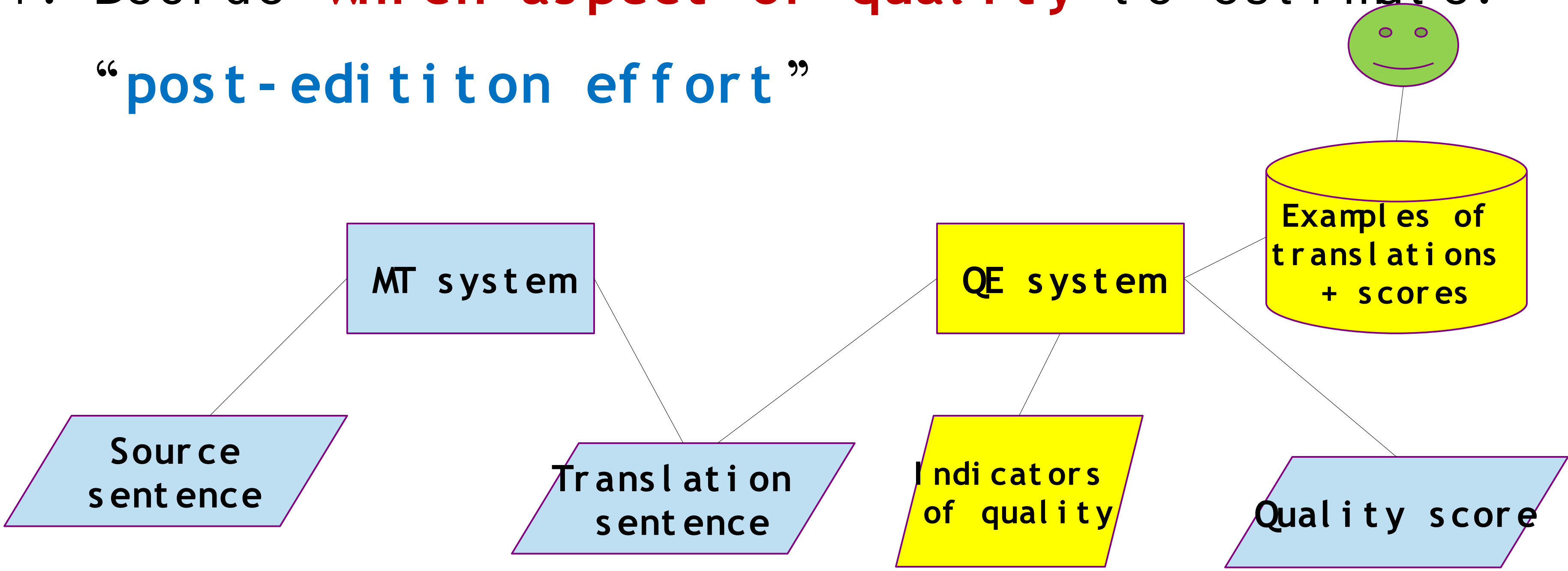
General approach

1. Decide **which aspect of quality** to estimate
2. Decide **how to represent** this aspect of quality
3. Collect **examples** of translations with different levels of quality
4. Identify and extract **indicators** that represent this quality
5. Apply an algorithm to induce a **model** to predict quality scores for new translations
6. **Evaluate** this model on new translations

General approach

1. Decide **which aspect of quality** to estimate:

   “**post-edititon effort**”



represent this quality

5. Apply an algorithm to induce a **model** to predict quality

How is quality defined?

1. **Good** vs **bad** translations: good for what? (Blatz et al. 2003)

2. **MT1** vs **MT 2**: is MT1 better than MT2. Yes, but is MT1 good enough? (Blatz et al. 2003; He et al., 2010)

3. **Perfect** vs **not perfect** translations: can we publish this translation as is? (Soricut and Echihabi 2010)

Define "quality" in terms of post-editing effort

4. Which translations are **good enough** for post-

How is quality defined?

What levels of quality can we expect from an MT system?

1. **Perfect**: no post-editing needed at all
2. **Good**: some post-editing needed, but **faster/easier** than translating from scratch
3. **Bad**: too much post-editing needed, faster/easier to translate from scratch

We expect the machine to estimate this well, but can humans do it well?

How is quality defined?

The court said that the rule was unjustified.

La cour a déclaré que la règle était injustifiée.

"I basically felt like I'd been exiled from my country and in forcing him to leave they'd also forced me to leave," she said.

"J'ai essentiellement ressenti si j'avais été exilé de mon pays et dans le forçant à quitter leur avais m'a aussi forcé de partir", dit-

How is quality defined?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle! …

Pour demain, et demain, et demain,
Creeps dans cette petite cadence de jour en jour,
Pour la dernière syllabe du temps enregistré;
Et tous nos hiers ont éclairé les fous
Le chemin de la mort poussiéreuse. Dehors, dehors, bougie bref! …

How do humans perform?

umans are good at identifying **perfect** translations, as well as terribly **bad** translations

ut medium quality translations are more difficult: “**good enough**” depends on the translator

- **Very experienced** translators: may prefer only close to perfect translations
- **Less experienced** translators: may benefit from

How do QE systems perform?

- **Humans:** agreement on **en-es Europarl**: 85% (prof., 2 an.)
- **Humans:** agreement on **en-pt subtitles** of TV series: 850 sentences (non prof, 3 an.)
  - 351 cases (41%) have **full** agreement
  - 445 cases (52%) have **partial** agreement
  - 54 cases (7%) have **null** agreement
- Agreement by s

Score	Full	Partial
4	59%	41%
3	35%	65%
2	23%	77%
1	50%	50%

How do QE systems perform?

implify the task, if we know how experienced the translator is: binary problem -> good

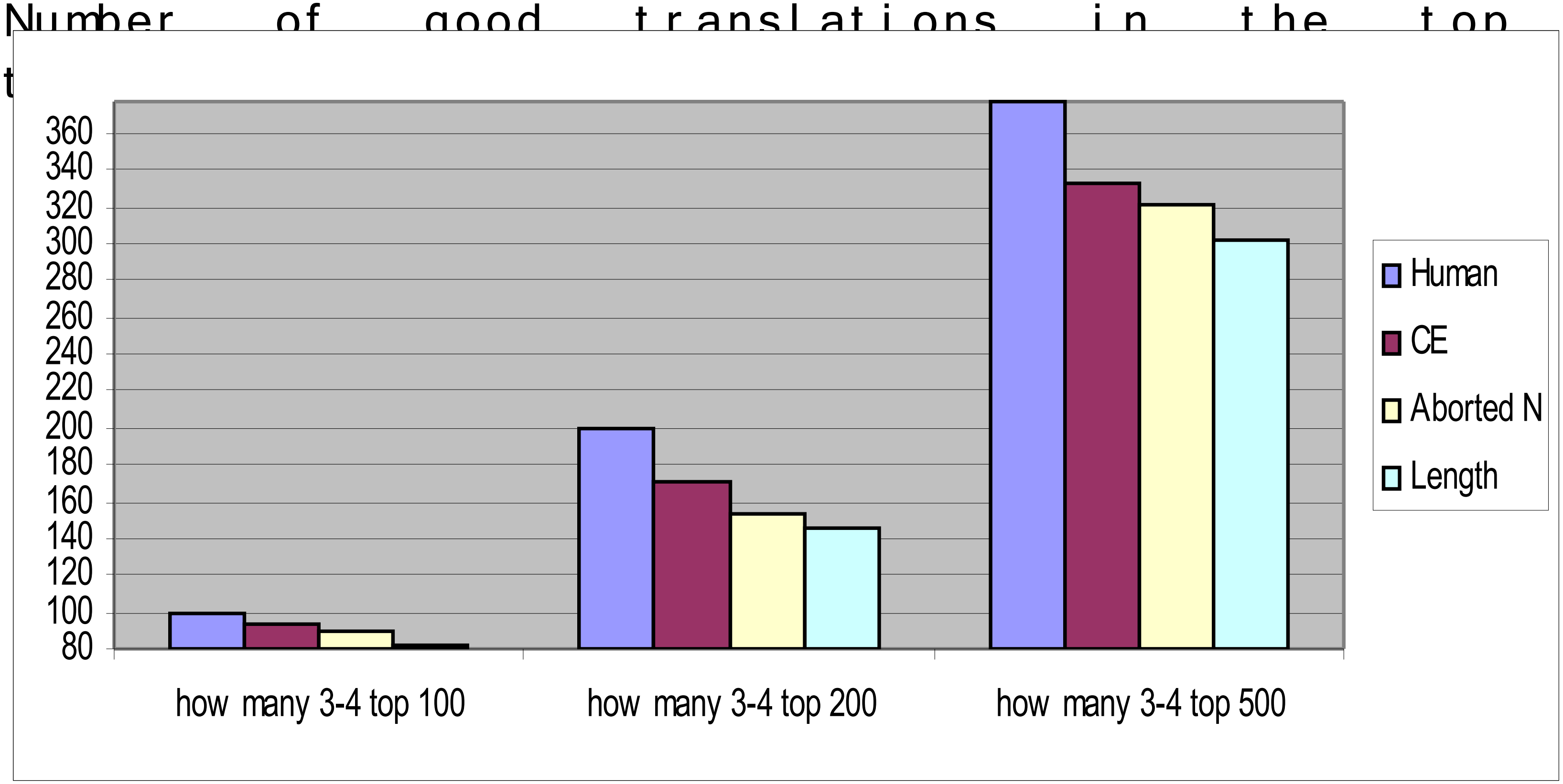
Languages	MT system	Accuracy	Most frequent score	Sentence length
en-es	MT1	70%	52%	36%
en-es	MT2	77%	74%	21%
en-es	MT3	66%	57%	30%
en-es	MT4	94%	94%	70%

How do QE systems perform?

- Evaluation in terms of **classification accuracy → clear**
  - Upper bound =100%
  - 50% = we are selecting 50% of the bad cases as good / of the good cases as bad
- **Is ~70% accuracy enough**?
- A different perspective: **precision/recall** by category:
  - How many bad translations the system says are good **(false rate)**
  - How many good the system says are bad **(miss rate)**

How do QE systems perform?

- Selecting only good translations: [3-4] (en-es)
  - Number of good translations in the top *n*
    t



Are 2/4 discrete scores enough?

- We want to estimate: **1,2** or **1, 2, 3, 4**
- It’s like saying you can get, from a TM:
  - Only **0% match** or **100% match**
  - Or the following (fuzzy) match levels: **0%, 50%, 75%, 100%**
- Isn’t there anything in between?

Estimate a continuum: a real number in [1,4]

Estimating a continuous score

- **English-Spanish Europarl** data
  - 4 SMT systems, 4 sets of **4,000 translations**
- Quality score: **1-4**

1: requires complete retranslation	2: a lot of post-editing needed, but quicker than retranslation
3: a little post-editing needed	4: fit for purpose

Languages	MT System	Error
en-es	MT1	0.653
en-es	MT2	0.718
en-es	MT3	0.706
en-es	MT4	0.603

Is a number in [1,4] informative?

an we see this number as a fuzzy match level?

- Not really... How much work to do on a 3.2 translation?

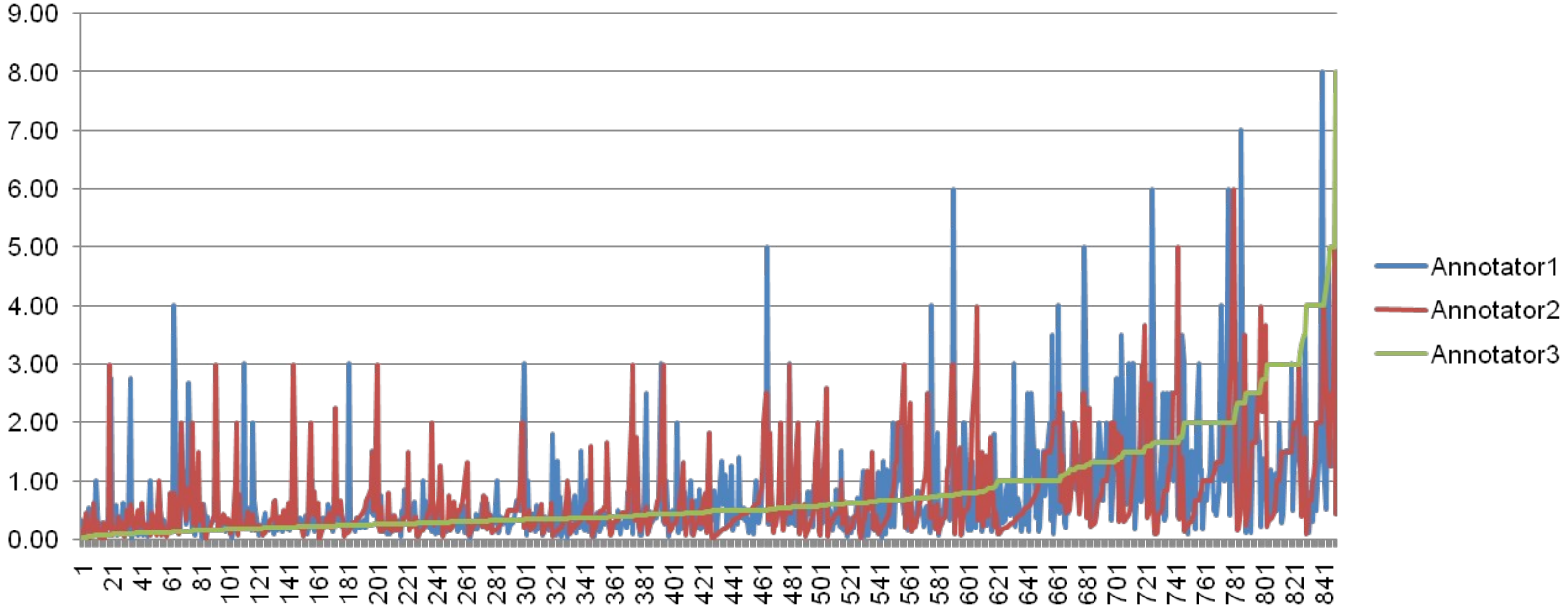
ry more objective ways of representing quality:

$$\text{HTER} = \frac{\#\ \text{edits}}{\#\ \text{words in post-edited version}}$$

- Edit distance (HTER): distance (in [0,1]) between original MT and post-edited version. What is the proportion of edits (words) will I have to perform to correct

Is a number in [1,4] informative?

- **Time** : how many seconds will it take to post-edit this sentence?
  - **Time varies** considerably from annotator to annotator



This annotation is **cheap** and **easy** to obtain if translators already post-edit MT

Other ways of representing quality

- English-Spanish, French-English news articles
- 1,500-2,500 translations
- Quality scores:
  - Score1 = HTER
  - Score2 = [1-4]
  - Score3 = time
- Annotation **tool** to collect data from translators

Other ways of representing quality

- Results
  - Each model trained on examples from a **single translator**

Dataset		Error ↓
fr-en	Distance	0.16
	[1-4]	0.66
	Time	0.65
en-es	Distance	0.18
	[1-4]	0.55
	Time	1.97

Other ways of representing quality

- So we are **almost** happy:
  - We can estimate an aspect of quality that is clear and objective (time, distance) ν

- But do these error metrics say something about how good the QE model is? Or which model is better? ✗

Evaluation by ranking

ank translations by their QE scores (best first)

ased on the quality of the MT system for a small development data, find the percentage of "**good enough**" translations, using any annotation scheme. E.g. 30% of the translation are good

easure improvement of top 30% according to QE scores:

- Compare **average quality of full dataset**

Evaluation by ranking

Languages	Delta [1-4] ↑	Delta Distance ↓ [0,1]	Delta Time ↓ (sec/word)
fr-en (70% good)	0.07	-0.02	-0.11
en-es (40% good)	0.20	-0.06	-0.19

Languages	Delta [1-4] ↑	Delta Distance ↓ [0,1]	Delta Time ↓ (sec/word)
fr-en	0.16	-0.04	-0.20
en-es	0.15	-0.04	-0.26

25%, 50% and 75%

Extrinsic evaluation by ranking

easure **post-editing time** to correct **top 30% translations** selected according to QE scores

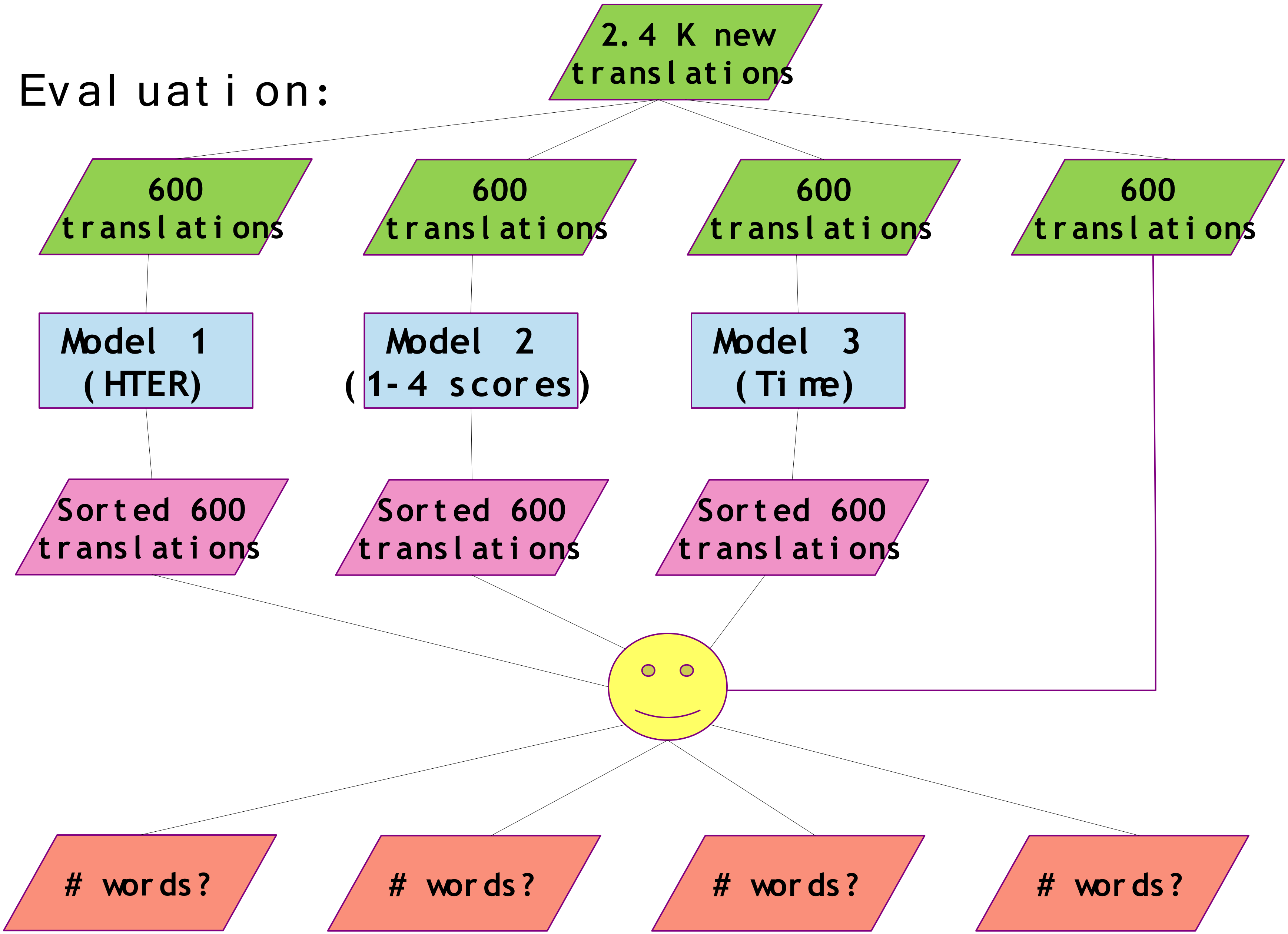
- Compare it against post-editing time of **randomly selected 30% translations**

f can't decide on the %, measure **number of words that can be post-edited in a fixed amount of time** from best to worse translations ranked according to QE model

- Compare it against number of words post-

Extrinsic evaluation by ranking

- Evaluation:



2.4 K new translations

600 translations

600 translations

600 translations

600 translations

Model 1 (HTER)

Model 2 (1-4 scores)

Model 3 (Time)

Sorted 600 translations

Sorted 600 translations

Sorted 600 translations

# words?

# words?

# words?

# words?

Extrinsic evaluation by ranking

- Post-editing in 1 hour:

MT System / Dataset	Words/second
S6 fr-en HTER (0-1)	0.96
S6 fr-en [1-4]	0.91
S6 fr-en time (sec/word)	**1.09**

MT System / Dataset	Words/second
S7 en-es HTER (0-1)	0.41
S7 en-es [1-4]	0.43
S7 en-es time (sec/word)	**0.57**
S7 en-es no CE	**0.32**

Extrinsic evaluation by ranking

umming up:

- The aspect of quality we estimate is clear (time, distance)
- The number ng-based (esp. extrinsic something about how good a

How about other users?

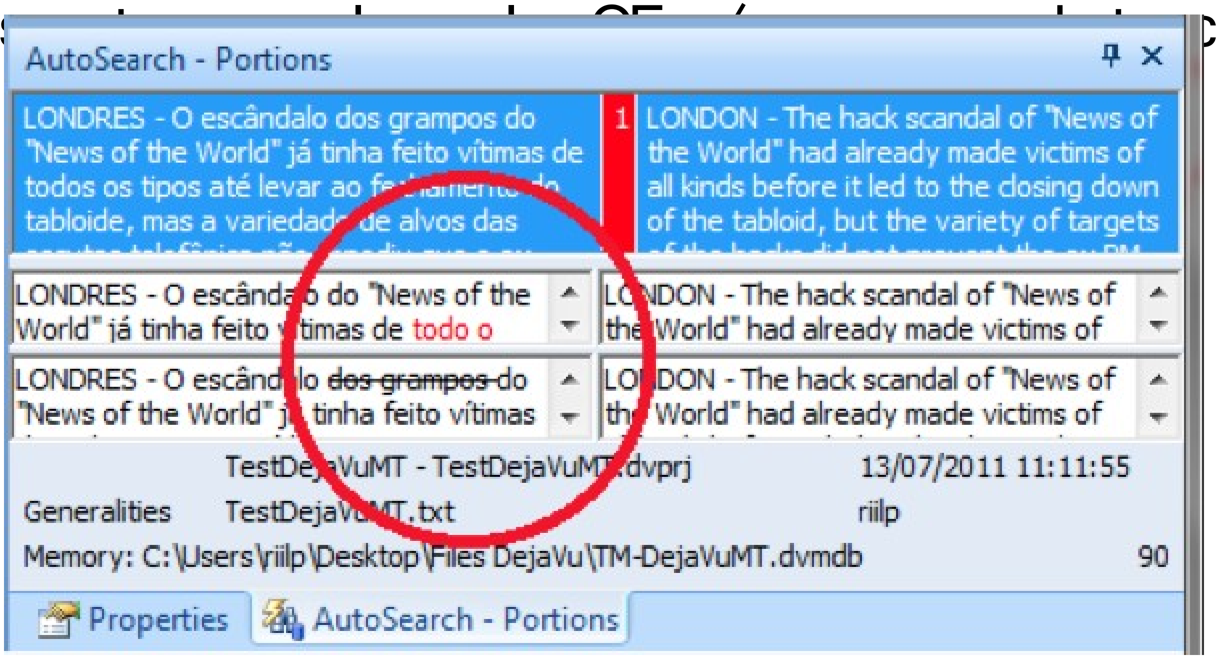
- Post-editing time/distance/[1-4] scores have a good (pearson) correlation:
  - **Distance** and **[1-4]** = **0.75 - 0.82**
  - **Time** and **[1-4]** = **0.50 - 0.60**

  (the smaller values are when scores are given by different translators)

- If we correlate post-editing **time/distance** and [1-4] scores reflecting **adequacy** (not post-editing effort)
  - **Distance** and **[1-4] Adequacy** = **0.55**
  - **Time** and **[1-4] Adequacy** = **0.40**

Is this enough?

- Is an accurate QE system at the sentence level enough?
- QE should also indicate, for sentences that are not perfect, **what the bad parts are**
  - ◆ Sub- ction in transl



LONDRES - O escândalo dos grampos do "News of the World" já tinha feito vítimas de todos os tipos até levar ao fechamento do tabloide, mas a variedade de alvos das	1	LONDON - The hack scandal of "News of the World" had already made victims of all kinds before it led to the closing down of the tabloid, but the variety of targets
LONDRES - O escândalo do "News of the World" já tinha feito vítimas de todo o		LONDON - The hack scandal of "News of the World" had already made victims of
LONDRES - O escândalo ~~dos grampos~~ do "News of the World" já tinha feito vítimas		LONDON - The hack scandal of "News of the World" had already made victims of

TestDejaVuMT - TestDejaVuMT.dvprj 13/07/2011 11:11:55

Generalities TestDejaVuMT.txt riilp

Memory: C:\Users\riilp\Desktop\Files DejaVu\TM-DejaVuMT.dvmdb 90

(Xiong et al., 2010): Link grammar: mostly words

Conclusions

- It is possible to estimate the quality of MT systems with respect to **post-editing needs**

- Measuring and estimating post-editing **time** seems to be the best way to **build** and **evaluate** QE systems
  - **Translator-dependent** measure: build a model per translator or project the time differences
  - Extrinsic evaluation using time is **expensive**, not feasible to compare many QE systems
  - Alternative: **intrinsic ranking-based**

Conclusions

- QE is a relatively **new area**

- It has a great potential to make MT more **useful** to end-users:
  - Translation: minimize post-editing time, allow for fair pricing models
  - Localization: keep the “brand” of the product/company
  - Gisting: avoid misunderstandings
  - Dissemination of large amounts of content, e.g.: user reviews

Advertisement:

- **Shared task on QE**
  - Most likely with **WMT** at NAACL, June 2012
  - **Sentence-level**: classification, regression and ranking

- We will provide:
  - Training sets annotated for quality
  - Baseline feature sets
  - Baseline systems to extract features
  - Test sets annotated for quality

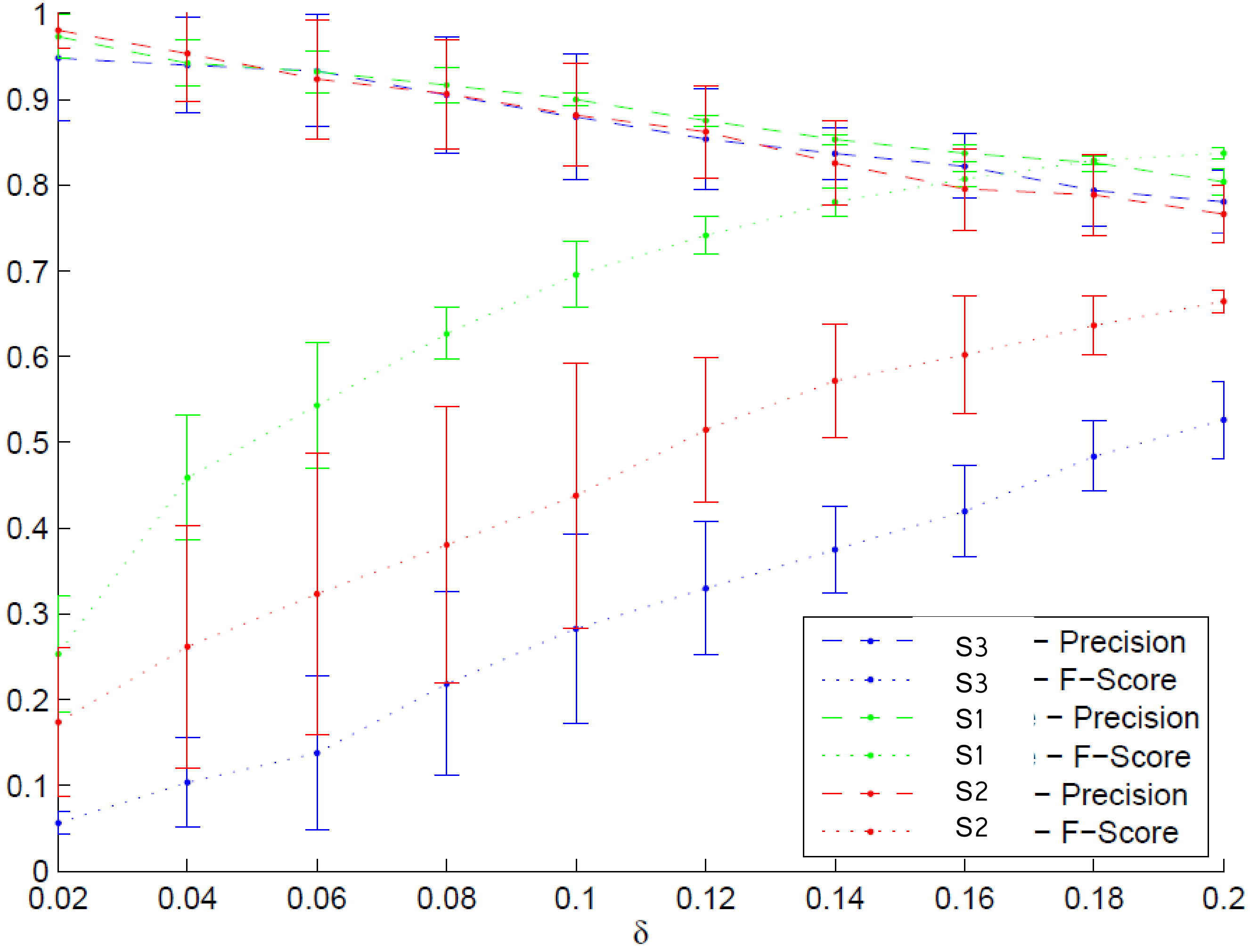
Questions?

Lucia Specia

l.specia@wlv.ac.uk

En-Es Europarl - [1-4]

- Regression + Confidence Machines to define the splitting point according to expected conf



$\tau = 3$

(MT Summit 2009)

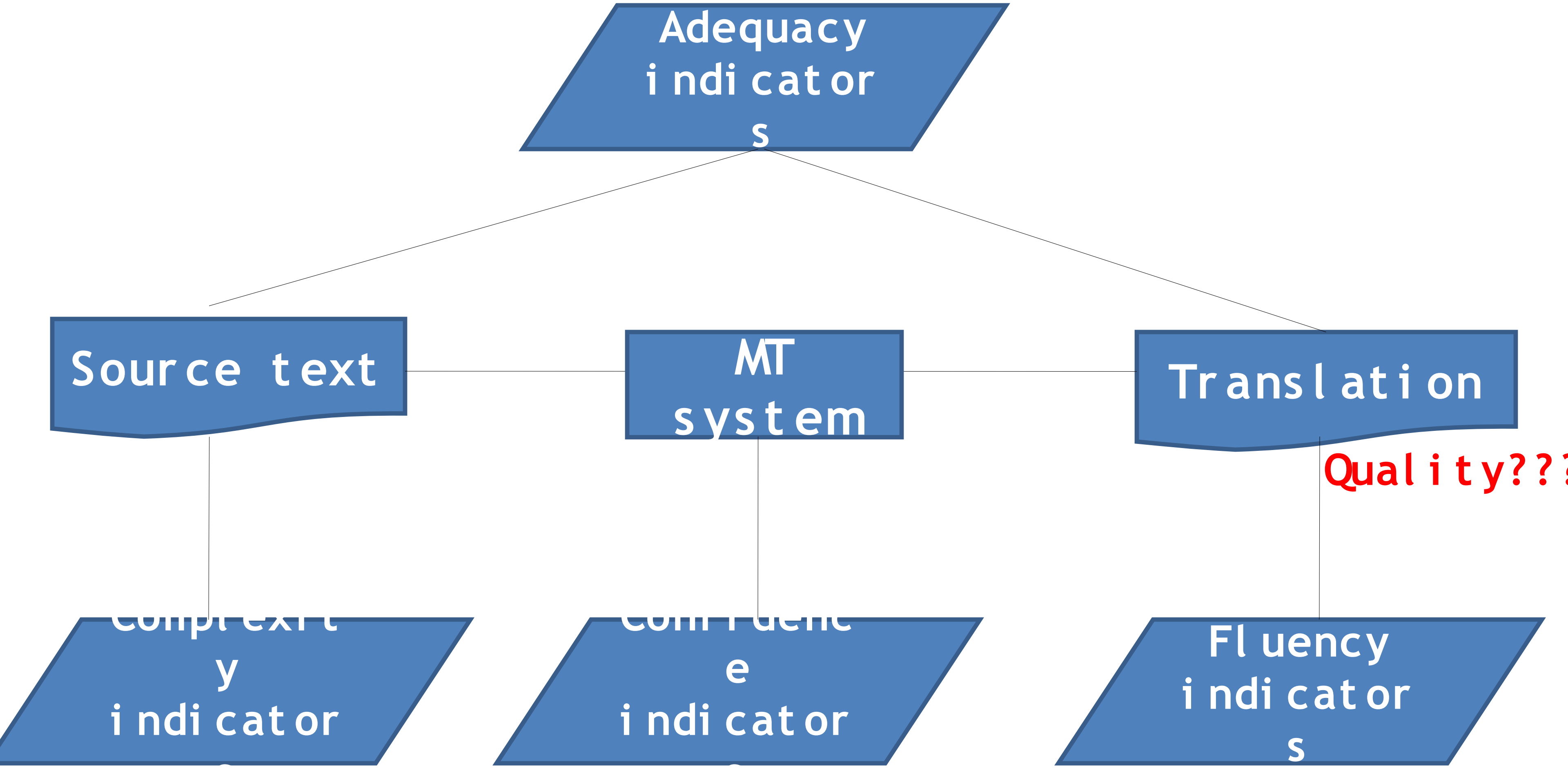
En-Es Europarl - [1- 4]

- QE score x MT metrics: Pearson's correlation across datasets produced by different MT systems:

Test set	Training set	Pearson QE and human
S3 en-es	S1 en-es	0.478
	S2 en-es	0.517
	S3 en-es	0.542
	S4 en-es	0.423
S2 en-es	S1 en-es	0.531
	S2 en-es	0.562
	S3 en-es	0.547
	S4 en-es	0.442

(Journal of MT 2010)

Features



- Shallow vs linguistically motivated
- MT system-dependent vs independent

Source features

- Source sentence length
- Language model of source
- Average number of possible translations per source word
- % of n-grams belonging to different frequency quartiles of the source side of the parallel corpus
- Average source word length
- ...

Target features

- ❑Target sentence length
- ❑Language model of target
- ❑Proportion of untranslated words
- ❑Grammar checking
- ❑Mismatching opening/closing brackets, quotation symbols
- ❑Coherence of the target sentence
- ❑…

MT features (confidence)

- SMT model global score and internal features
  - Distortion count, phrase probability, ...
- % search nodes aborted, pruned, recombined …
- Language model using n-best list as corpus
- Distance to centre hypothesis in the n-best list
- Relative frequency of the words in the translation in the n-best list
- Ratio of SMT model score of the top translation to the sum of the scores of all hypothesis in the n-best list
- …

Source-target features

- Ratio between source and target sentence lengths
- Punctuation checking (target vs source)
- Correct translation of pronouns
- Matching of phrase/POS tags
- Matching of dependency relations
- Matching of named entities
- Matching of semantic role labels
- Alignment of these and other linguistic markers
- ...

MT system selection

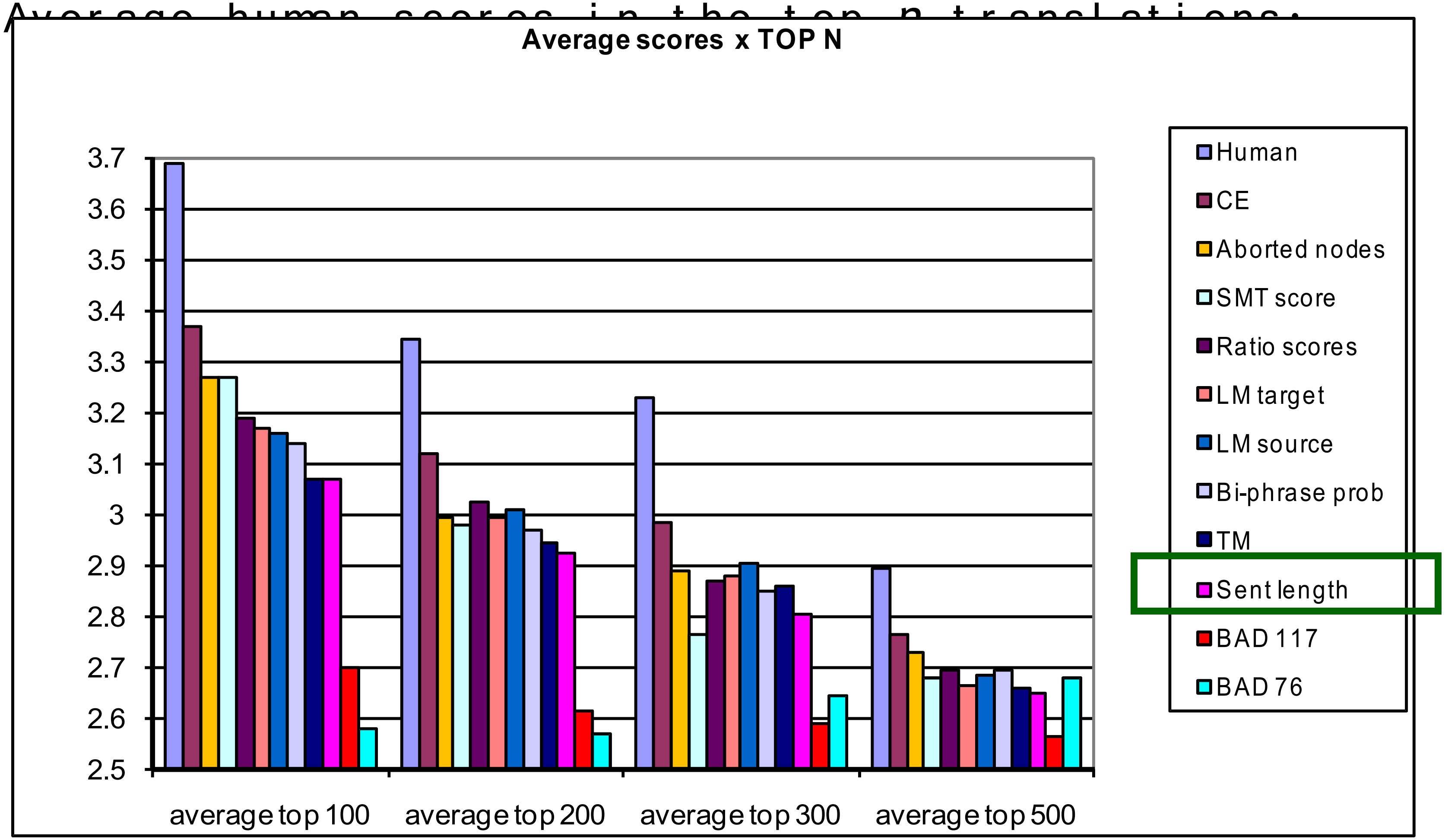
- **Approach**:
  - **Train** QE models for each MT system (**individually**)
  - Use all MT systems to **translate** each input segment
  - **Estimate** the QE score for each alternative translation
  - **Select** the translation with the highest CE score

- Experiments:
  - **En-Es Europarl [1-4] datasets**, 4 MT systems

- Results:

How do QE systems perform?

- Selecting only good translations: [3-4] (en-es)
  - Average human scores in the top n translations:


Average scores x TOP N
average top 100
average top 200
average top 300
average top 500
Human
CE
Aborted nodes
SMT score
Ratio scores
LM target
LM source
Bi-phrase prob
TM
Sent length
BAD 117
BAD 76