# Exploiting Objective Annotations for Measuring Translation Post-editing Effort

EAMT 2011

Lucia Specia

University of Wolverhampton
l.specia@wlv.ac.uk

30 May 2011

# Outline

# Introduction

**MT**: Events of a magnitude unprecedented Mongols claiming their rights have occurred last week in this autonomous region, according to the Information Centre on Human Rights in Shouth Mongolia, an organization based in the States U.S., where universities and public spaces open air were banned from several cities, fearing the power to Beijing more than any protest rallies in the spirit of movements which have stirred recent months the world Arabic.

**SRC**: Des manifestations d'une ampleur sans précédent de Mongols réclamant le respect de leurs droits se sont produites la semaine dernière dans cette région autonome, selon le Centre d'information sur les droits de l'homme en Mongolie du Sud, une organisation installée aux Etats-Unis, où des universités et des espaces publics en plein air étaient interdits d'accès dans plusieurs villes, le pouvoir à Pékin redoutant plus que tout des rassemblements de protestation dans l'esprit des mouvements qui ont agité ces derniers mois des pays du monde arabe.

# Introduction

**MT**: Events of a magnitude unprecedented Mongols claiming their rights have occurred last week in this autonomous region, according to the Information Centre on Human Rights in Shouth Mongolia, an organization based in the States U.S., where universities and public spaces open air were banned from several cities, fearing the power to Beijing more than any protest rallies in the spirit of movements which have stirred recent months the world Arabic.

**SRC**: Des manifestations d'une ampleur sans précédent de Mongols réclamant le respect de leurs droits se sont produites la semaine dernière dans cette région autonome, selon le Centre d'information sur les droits de l'homme en Mongolie du Sud, une organisation installée aux Etats-Unis, où des universités et des espaces publics en plein air étaient interdits d'accès dans plusieurs villes, le pouvoir à Pékin redoutant plus que tout des rassemblements de protestation dans l'esprit des mouvements qui ont agité ces derniers mois des pays du monde arabe.

# Introduction

- **Post-editing of MT output** is a common practice for many human translators
- However, certain translated segments may require more post-editing than others:
  - It may be **faster** to translate some segments from scratch
  - Filtering out bad translations can prevent **translators frustration**
  - Distinguishing bad from good translations allows fairer **cost schemes**
- The problem of distinguishing bad from good translations is addressed by metrics of **Quality Estimation (QE)**

# Introduction

- **Post-editing of MT output** is a common practice for many human translators
- However, certain translated segments may require more post-editing than others:
  - It may be **faster** to translate some segments from scratch
  - Filtering out bad translations can prevent **translators frustration**
  - Distinguishing bad from good translations allows fairer **cost schemes**
- The problem of distinguishing bad from good translations is addressed by metrics of **Quality Estimation (QE)**

# Introduction

- **Post-editing of MT output** is a common practice for many human translators
- However, certain translated segments may require more post-editing than others:
  - It may be **faster** to translate some segments from scratch
  - Filtering out bad translations can prevent **translators frustration**
  - Distinguishing bad from good translations allows fairer **cost schemes**
- The problem of distinguishing bad from good translations is addressed by metrics of **Quality Estimation (QE)**

# Introduction

- **Post-editing of MT output** is a common practice for many human translators
- However, certain translated segments may require more post-editing than others:
  - It may be **faster** to translate some segments from scratch
  - Filtering out bad translations can prevent **translators frustration**
  - Distinguishing bad from good translations allows fairer **cost schemes**
- The problem of distinguishing bad from good translations is addressed by metrics of **Quality Estimation (QE)**

# Introduction

- **Post-editing of MT output** is a common practice for many human translators
- However, certain translated segments may require more post-editing than others:
  - It may be **faster** to translate some segments from scratch
  - Filtering out bad translations can prevent **translators frustration**
  - Distinguishing bad from good translations allows fairer **cost schemes**
- The problem of distinguishing bad from good translations is addressed by metrics of **Quality Estimation (QE)**

# Outline

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Quality Estimation

- **Features** extracted from:
  - MT output
  - Source text
  - Monolingual corpora: source or/and target
  - Bilingual corpora
  - MT system (CE)
- **Annotations** reflecting translation quality
- Train a machine **learning algorithm** to produce a model for a certain:
  - Language pair
  - MT system
  - **(Ideally) Text domain & genre**
  - **(Ideally) Human translator**

# Outline

# Related Work

- CE metrics may provide a score for each:
  - **word or phrase** [GF03, UN05, KN06]
  - **sentence** [BFF+04, Qui04, STC+09, SRT10, HMvGW10, SF10]
  - **document** [SE10]
- Quality **annotation** can be derived using:
  - **Automatic MT evaluation metrics** [BFF+04]
  - **Human annotation**: proved better [Qui04, STC+09]
- **Human annotation** can be expensive and subjective
  - No previous studies comparing different forms of human annotation

# Related Work

- CE metrics may provide a score for each:
  - **word or phrase** [GF03, UN05, KN06]
  - **sentence** [BFF+04, Qui04, STC+09, SRT10, HMvGW10, SF10]
  - **document** [SE10]
- Quality **annotation** can be derived using:
  - **Automatic MT evaluation metrics** [BFF+04]
  - **Human annotation**: proved better [Qui04, STC+09]
- **Human annotation** can be expensive and subjective
  - No previous studies comparing different forms of human annotation

# Related Work

- CE metrics may provide a score for each:
  - **word or phrase** [GF03, UN05, KN06]
  - **sentence**
    [BFF+04, Qui04, STC+09, SRT10, HMvGW10, SF10]
  - **document** [SE10]
- Quality **annotation** can be derived using:
  - **Automatic MT evaluation metrics** [BFF+04]
  - **Human annotation**: proved better [Qui04, STC+09]
- **Human annotation** can be expensive and subjective
  - No previous studies comparing different forms of human annotation

# Outline

# Goals

- Measure the **post-editing time** for unseen **sentence translations** predicted as "good quality" according to QE models learnt based on different types of human annotation:
  - Absolute scores reflecting post-editing effort
  - Edit distance between automatic and post-edited translations (HTER)
  - Post-editing time
- Show that using such QE models to select a subset of translations for post-editing can **speed up post-editing tasks**

# Goals

- Measure the **post-editing time** for unseen **sentence translations** predicted as "good quality" according to QE models learnt based on different types of human annotation:
  - Absolute scores reflecting post-editing effort
  - Edit distance between automatic and post-edited translations (HTER)
  - Post-editing time
- Show that using such QE models to select a subset of translations for post-editing can **speed up post-editing tasks**

# Goals

- Measure the **post-editing time** for unseen **sentence translations** predicted as "good quality" according to QE models learnt based on different types of human annotation:
  - Absolute scores reflecting post-editing effort
  - Edit distance between automatic and post-edited translations (HTER)
  - Post-editing time
- Show that using such QE models to select a subset of translations for post-editing can **speed up post-editing tasks**

# Goals

- Measure the **post-editing time** for unseen **sentence translations** predicted as "good quality" according to QE models learnt based on different types of human annotation:
  - Absolute scores reflecting post-editing effort
  - Edit distance between automatic and post-edited translations (HTER)
  - Post-editing time
- Show that using such QE models to select a subset of translations for post-editing can **speed up post-editing tasks**

# Goals

- Measure the **post-editing time** for unseen **sentence translations** predicted as “good quality” according to QE models learnt based on different types of human annotation:
  - Absolute scores reflecting post-editing effort
  - Edit distance between automatic and post-edited translations (HTER)
  - Post-editing time
- Show that using such QE models to select a subset of translations for post-editing can **speed up post-editing tasks**

# Goals

- Hypothesis is that **simpler**, **cheaper**, **more transparent** and **more objective** annotations can have a more straightforward interpretation for post-editing purposes

# Outline

1. Introduction
2. Quality Estimation
3. Related Work
4. Goals
5. Datasets
6. Results
7. Conclusions

# Datasets

- Datasets collected using *news* source sentences from WMT's development and test sets

# Datasets

- Datasets collected using *news* source sentences from WMT's development and test sets
- Translations produced using a standard phrase-based SMT (Moses):

# Datasets

- Datasets collected using *news* source sentences from WMT's development and test sets
- Translations produced using a standard phrase-based SMT (Moses):
  - **fr-en news-test2009**: 2,525 French news sentences and their translations into English (BLEU = 0.2447)
  - **en-es news-test2010**: 1,000 English news sentences and their translations into Spanish (BLEU = 0.2830)

# Datasets

- **Post-editing tool** similar interface to TM tools: shows the **source** sentence and the machine **translation** for post-editing

# Datasets

- **Post-editing tool** similar interface to TM tools: shows the **source** sentence and the machine **translation** for post-editing
- Translators instructed to perform the **minimum number of editions** necessary to make the translation ready for publishing

# Datasets

- **Post-editing tool** similar interface to TM tools: shows the **source** sentence and the machine **translation** for post-editing
- Translators instructed to perform the **minimum number of editions** necessary to make the translation ready for publishing
- **Post-editing time** is measured on a sentence-basis

# Datasets

- **Post-editing tool** similar interface to TM tools: shows the **source** sentence and the machine **translation** for post-editing
- Translators instructed to perform the **minimum number of editions** necessary to make the translation ready for publishing
- **Post-editing time** is measured on a sentence-basis
- Translators also scored the original translation according to its **post-editing effort**:

# Datasets - Annotation

- **Post-editing effort score** (*effort*): a discrete score:

# Datasets - Annotation

- **Post-editing effort score** (*effort*): a discrete score:
  - 1 = requires complete retranslation
  - 2 = post editing still quicker than retranslation
  - 3 = very little post editing needed
  - 4 = fit for purpose

# Datasets - Annotation

- **Post-editing effort score** (*effort*): a discrete score:
  - 1 = requires complete retranslation
  - 2 = post editing still quicker than retranslation
  - 3 = very little post editing needed
  - 4 = fit for purpose
- **Post-editing distance** (*HTER*): a continuous score in $[0, 1]$:

$$\text{HTER} = \frac{\#edits}{\#words\_postedited\_version}$$

# Datasets - Annotation

- **Post-editing effort score** (*effort*): a discrete score:
  - 1 = requires complete retranslation
  - 2 = post editing still quicker than retranslation
  - 3 = very little post editing needed
  - 4 = fit for purpose
- **Post-editing distance** (*HTER*): a continuous score in $[0, 1]$:

$$\text{HTER} = \frac{\#edits}{\#words\_postedited\_version}$$

- **Post-editing time** (*time*): average number of **seconds to post-edit each word** in the sentence

# Quality Estimation Framework

- Similar to that proposed by [SF10], with **SVM for regression**: epsilon-SVR algorithm with radial basis function kernel from the LIBSVM package [CL01], with the parameters $\gamma$, $\epsilon$ and *cost* optimized.
- 80 **shallow, MT system-independent features**:
  - source & target sentence lengths and their ratios
  - source & target sentence type/token ratio
  - average source word length
  - average number of occurrences of all target words within the target sentence
  - source & target sentence 3-gram LM probabilities and perplexities

# Quality Estimation Framework

- Similar to that proposed by [SF10], with **SVM for regression**: epsilon-SVR algorithm with radial basis function kernel from the LIBSVM package [CL01], with the parameters $\gamma$, $\epsilon$ and *cost* optimized.
- 80 **shallow, MT system-independent features**:
  - source & target sentence lengths and their ratios
  - source & target sentence type/token ratio
  - average source word length
  - average number of occurrences of all target words within the target sentence
  - source & target sentence 3-gram LM probabilities and perplexities

# Quality Estimation Framework

- percentage of 1 to 3-grams in the source sentence belonging to each frequency quartile of a source corpus
- average number of translations per source word in the sentence (given by GIZA++ tables), unweighted/weighted by the (inverse) frequency of words
- percentages of numbers, content- / non-content words in the source & target sentences
- number of mismatching opening/closing brackets and quotation marks in the target sentence
- percentages & number of mismatches of some superficial constructions between the source and target sentences: brackets, punctuation symbols, numbers

# Outline

# Results

## Average Human Scores

| Dataset | | Average Human Score |
|---|---|---|
| fr-en | *HTER* | 0.201 ↓ |
| | *effort* | 2.834 ↑ |
| | *time snt* | 24.095 ↓ |
| en-es | *HTER* | 0.349 ↓ |
| | *effort* | 2.441 ↑ |
| | *time snt* | 98.692 ↓ |

- Translators have different **level of experience**: en-es translator is more experienced
- Translators followed **different strategies**: fr-en translator read the source before the time measurement started

# Results
## Average Human Scores

| Dataset | | Average Human Score |
|---|---|---|
| fr-en | *HTER* | 0.201 ↓ |
| | *effort* | 2.834 ↑ |
| | *time snt* | 24.095 ↓ |
| en-es | *HTER* | 0.349 ↓ |
| | *effort* | 2.441 ↑ |
| | *time snt* | 98.692 ↓ |

- Translators have different **level of experience**: en-es translator is more experienced
- Translators followed **different strategies**: fr-en translator read the source before the time measurement started

# Results

## Average Human Scores

| Dataset | | Average Human Score |
|---|---|---|
| fr-en | *HTER* | 0.201 ↓ |
| | *effort* | 2.834 ↑ |
| | *time snt* | 24.095 ↓ |
| en-es | *HTER* | 0.349 ↓ |
| | *effort* | 2.441 ↑ |
| | *time snt* | 98.692 ↓ |

- Translators have different **level of experience**: en-es translator is more experienced
- Translators followed **different strategies**: fr-en translator read the source before the time measurement started

# Results
## Prediction Error and Correlation

- Spearman's **rank coefficient** with human scores

# Results

## Prediction Error and Correlation

- Spearman's **rank coefficient** with human scores
- **Root Mean Squared Error** (RMSE) for regression error

# Results

## Prediction Error and Correlation

- Spearman's **rank coefficient** with human scores
- **Root Mean Squared Error** (RMSE) for regression error
- 5-fold cross validation: training on 90% and test on 10%

# Results

## Prediction Error and Correlation

- Spearman's **rank coefficient** with human scores
- **Root Mean Squared Error** (RMSE) for regression error
- 5-fold cross validation: training on 90% and test on 10%

| **Dataset** | | **RMSE** $\downarrow$ | **Spearman** $\uparrow$ |
|---|---|---|---|
| fr-en | *HTER* | $0.155 \pm 0.011$ | $0.366 \pm 0.047$ |
| | *effort* | $0.662 \pm 0.022$ | $0.459 \pm 0.034$ |
| | *time* | $0.651 \pm 0.040$ | $0.455 \pm 0.052$ |
| en-es | *HTER* | $0.178 \pm 0.006$ | $0.281 \pm 0.102$ |
| | *effort* | $0.549 \pm 0.028$ | $0.367 \pm 0.096$ |
| | *time* | $1.970 \pm 0.250$ | $0.298 \pm 0.024$ |

# Results

## Task-based Evaluation

- **Goal**: measure number of words that can be post-edited in a fixed amount of time in translations selected according to each QE model

# Results
## Task-based Evaluation

- **Goal**: measure number of words that can be post-edited in a fixed amount of time in translations selected according to each QE model
- **Unseen sentences** with the same genre and domain translated using Moses:

# Results
## Task-based Evaluation

- **Goal**: measure number of words that can be post-edited in a fixed amount of time in translations selected according to each QE model
- **Unseen sentences** with the same genre and domain translated using Moses:
  - **fr-en news-test2010**: 2,489 French news sentences and their translations into English (BLEU = 0.2551)
  - **en-es news-test2009**: 2,525 English news sentences and their translations into Spanish (BLEU = 0.2428)

# Results
## Task-based Evaluation

- **Goal**: measure number of words that can be post-edited in a fixed amount of time in translations selected according to each QE model
- **Unseen sentences** with the same genre and domain translated using Moses:
  - **fr-en news-test2010**: 2,489 French news sentences and their translations into English (BLEU = 0.2551)
  - **en-es news-test2009**: 2,525 English news sentences and their translations into Spanish (BLEU = 0.2428)
- **Quality predictions** generated using the 3 variations of the QE models

# Results

## Task-based Evaluation

- Predicted scores can be used to directly **filter out bad quality translations**:
  - Setting a threshold on estimated scores: [STW+09], [HMvGW10]
- We evaluate the **ranking of translations** using QE scores from alternative models in order to answer:
  1. Which annotation type yields models that allow ranking sentences so that selecting the top ranked sentences can **maximize the number of words that can be post-edited per second**?
  2. Using such models to rank sentences and selecting the top ranked sentences, is it possible to **post-edit more words** as compared to post-editing sentences without any ranking in a given slot of time?

# Results
## Task-based Evaluation

- Predicted scores can be used to directly **filter out bad quality translations**:
  - Setting a threshold on estimated scores: [STW+09], [HMvGW10]
- We evaluate the **ranking of translations** using QE scores from alternative models in order to answer:
  1. Which annotation type yields models that allow ranking sentences so that selecting the top ranked sentences can **maximize the number of words that can be post-edited per second**?
  2. Using such models to rank sentences and selecting the top ranked sentences, is it possible to **post-edit more words** as compared to post-editing sentences without any ranking in a given slot of time?

# Results
## Task-based Evaluation

- Predicted scores can be used to directly **filter out bad quality translations**:
  - Setting a threshold on estimated scores: [STW$^+$09], [HMvGW10]
- We evaluate the **ranking of translations** using QE scores from alternative models in order to answer:
  1. Which annotation type yields models that allow ranking sentences so that selecting the top ranked sentences can **maximize the number of words that can be post-edited per second**?
  2. Using such models to rank sentences and selecting the top ranked sentences, is it possible to **post-edit more words** as compared to post-editing sentences without any ranking in a given slot of time?

# Results
## Task-based Evaluation

- Predicted scores can be used to directly **filter out bad quality translations**:
  - Setting a threshold on estimated scores: [STW+09], [HMvGW10]
- We evaluate the **ranking of translations** using QE scores from alternative models in order to answer:
  1. Which annotation type yields models that allow ranking sentences so that selecting the top ranked sentences can **maximize the number of words that can be post-edited per second**?
  2. Using such models to rank sentences and selecting the top ranked sentences, is it possible to **post-edit more words** as compared to post-editing sentences without any ranking in a given slot of time?

# Results
## Task-based Evaluation

- Predicted scores can be used to directly **filter out bad quality translations**:
  - Setting a threshold on estimated scores: [STW+09], [HMvGW10]
- We evaluate the **ranking of translations** using QE scores from alternative models in order to answer:
  1. Which annotation type yields models that allow ranking sentences so that selecting the top ranked sentences can **maximize the number of words that can be post-edited per second**?
  2. Using such models to rank sentences and selecting the top ranked sentences, is it possible to **post-edit more words** as compared to post-editing sentences without any ranking in a given slot of time?

# Results
## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results
## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results

## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results

## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results
## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results

## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results
## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results

## Task-based Evaluation

- **4 subsets of** 600 **translations** randomly selected from each unseen dataset
  - Translations in **3 subsets ranked using each QE model** so that the best translations appear first
  - Translations in **1 subset not ranked**
- Translators asked to post-edited as many sentences as possible in each of 4 "tasks" on different days:
  - **1 hour per task**
  - Tasks order:
    - T1: 600 MT sentences sorted acc. to *HTER* model
    - T2: 600 MT sentences sorted acc. to *effort* model
    - T3: 600 MT sentences sorted acc. to *time* model
    - T4: 600 MT sentences without any sorting
- Variation: *effort* in en-es datasets: 43% "good" (4-3), 57% "bad" (1-2)

# Results
## Task-based Evaluation

| Dataset | | Sentences/h ↑ | Words/s ↑ |
|---|---|---|---|
| fr-en | T1: *HTER* | 65 | 0.96 |
| | T2: *effort* | 97 | 0.91 |
| | T3: *time* | 82 | 1.09 |
| | T4: unsorted | 55 | 0.75 |
| en-es | T1: *HTER* | 38 | 0.41 |
| | T2: *effort* | 71 | 0.43 |
| | T3: *time* | 69 | 0.57 |
| | T4: unsorted | 33 | 0.32 |

- **Post-editing only top translations** acc. to any QE model: more words post-edited per second than post-editing any translation
- **Best rate obtained with time**: both **fr-en** and **en-es**

# Results

## Task-based Evaluation

| Dataset | | Sentences/h ↑ | Words/s ↑ |
|---|---|---|---|
| fr-en | T1: *HTER* | 65 | 0.96 |
| | T2: *effort* | 97 | 0.91 |
| | T3: *time* | 82 | 1.09 |
| | T4: unsorted | 55 | 0.75 |
| en-es | T1: *HTER* | 38 | 0.41 |
| | T2: *effort* | 71 | 0.43 |
| | T3: *time* | 69 | 0.57 |
| | T4: unsorted | 33 | 0.32 |

- **Post-editing only top translations** acc. to any QE model: more words post-edited per second than post-editing any translation
- **Best rate obtained with time**: both **fr-en** and **en-es**

# Results
## Task-based Evaluation

| Dataset | | Sentences/h ↑ | Words/s ↑ |
|---|---|---|---|
| fr-en | T1: *HTER* | 65 | 0.96 |
| | T2: *effort* | 97 | 0.91 |
| | T3: *time* | 82 | 1.09 |
| | T4: unsorted | 55 | 0.75 |
| en-es | T1: *HTER* | 38 | 0.41 |
| | T2: *effort* | 71 | 0.43 |
| | T3: *time* | 69 | 0.57 |
| | T4: unsorted | 33 | 0.32 |

- **Post-editing only top translations** acc. to any QE model: more words post-edited per second than post-editing any translation
- **Best rate obtained with time**: both **fr-en** and **en-es**

# Outline

# Conclusions

- We have presented experiments with alternative ways of annotating translation quality for building QE models
- Explicit and subjective annotations used in previous work, **post-editing effort**, are worse than simpler and more objective metrics, in particular **time**
- These can be obtained as a **by-product** of having **humans post-editing** a reasonably **small number** of translations
- **Translators are different**: QE model for each human translator (MT system, language pair)

# Conclusions

- We have presented experiments with alternative ways of annotating translation quality for building QE models
- Explicit and subjective annotations used in previous work, **post-editing effort**, are worse than simpler and more objective metrics, in particular **time**
- These can be obtained as a **by-product** of having **humans post-editing** a reasonably **small number** of translations
- **Translators are different**: QE model for each human translator (MT system, language pair)

# Conclusions

- We have presented experiments with alternative ways of annotating translation quality for building QE models
- Explicit and subjective annotations used in previous work, **post-editing effort**, are worse than simpler and more objective metrics, in particular **time**
- These can be obtained as a **by-product** of having **humans post-editing** a reasonably **small number** of translations
- **Translators are different**: QE model for each human translator (MT system, language pair)

# Conclusions

- We have presented experiments with alternative ways of annotating translation quality for building QE models
- Explicit and subjective annotations used in previous work, **post-editing effort**, are worse than simpler and more objective metrics, in particular **time**
- These can be obtained as a **by-product** of having **humans post-editing** a reasonably **small number** of translations
- **Translators are different**: QE model for each human translator (MT system, language pair)

# Conclusions

- In real world scenarios translators would have to **translate all sentences** - not only the top ranked ones
- A reliable model can help distinguishing sentences that are worth post-editing from those that should be translated in order to:
  - **Increase productivity** by preventing translators from spending time reading bad quality translations
  - Minimize **translators' frustration** with trying to post-edit bad quality translations
- Datasets are **available for download**

# Conclusions

- In real world scenarios translators would have to **translate all sentences** - not only the top ranked ones
- A reliable model can help distinguishing sentences that are worth post-editing from those that should be translated in order to:
  - **Increase productivity** by preventing translators from spending time reading bad quality translations
  - Minimize **translators' frustration** with trying to post-edit bad quality translations
- Datasets are **available for download**

# Conclusions

- In real world scenarios translators would have to **translate all sentences** - not only the top ranked ones
- A reliable model can help distinguishing sentences that are worth post-editing from those that should be translated in order to:
  - **Increase productivity** by preventing translators from spending time reading bad quality translations
  - Minimize **translators' frustration** with trying to post-edit bad quality translations
- Datasets are **available for download**

# Conclusions

- In real world scenarios translators would have to **translate all sentences** - not only the top ranked ones
- A reliable model can help distinguishing sentences that are worth post-editing from those that should be translated in order to:
  - **Increase productivity** by preventing translators from spending time reading bad quality translations
  - Minimize **translators' frustration** with trying to post-edit bad quality translations
- Datasets are **available for download**

# Conclusions

- In real world scenarios translators would have to **translate all sentences** - not only the top ranked ones
- A reliable model can help distinguishing sentences that are worth post-editing from those that should be translated in order to:
  - **Increase productivity** by preventing translators from spending time reading bad quality translations
  - Minimize **translators' frustration** with trying to post-edit bad quality translations
- Datasets are **available for download**

# Conclusions

- In real world scenarios translators would have to **translate all sentences** - not only the top ranked ones
- A reliable model can help distinguishing sentences that are worth post-editing from those that should be translated in order to:
    - **Increase productivity** by preventing translators from spending time reading bad quality translations
    - Minimize **translators' frustration** with trying to post-edit bad quality translations
- Datasets are **available for download**

# Future work

- Combine these algorithms with techniques to establish **thresholds on the predicted scores**
- Design a **post-editing tool** that can incorporate quality predictions for translations from different MT/TM systems
- Analyze **changes in the behavior of translators** as they gain more experience with the task of post-editing, especially wrt post-editing time
- Use crowdsourcing mechanisms to include **other language pairs** and **multiple post-editors and reviewers**

# Future work

- Combine these algorithms with techniques to establish **thresholds on the predicted scores**
- Design a **post-editing tool** that can incorporate quality predictions for translations from different MT/TM systems
- Analyze **changes in the behavior of translators** as they gain more experience with the task of post-editing, especially wrt post-editing time
- Use crowdsourcing mechanisms to include **other language pairs** and **multiple post-editors and reviewers**

# Future work

- Combine these algorithms with techniques to establish **thresholds on the predicted scores**
- Design a **post-editing tool** that can incorporate quality predictions for translations from different MT/TM systems
- Analyze **changes in the behavior of translators** as they gain more experience with the task of post-editing, especially wrt post-editing time
- Use crowdsourcing mechanisms to include **other language pairs** and **multiple post-editors and reviewers**

# Future work

- Combine these algorithms with techniques to establish **thresholds on the predicted scores**
- Design a **post-editing tool** that can incorporate quality predictions for translations from different MT/TM systems
- Analyze **changes in the behavior of translators** as they gain more experience with the task of post-editing, especially wrt post-editing time
- Use crowdsourcing mechanisms to include **other language pairs** and **multiple post-editors and reviewers**

# Exploiting Objective Annotations for Measuring Translation Post-editing Effort

EAMT 2011

Lucia Specia

University of Wolverhampton
l.specia@wlv.ac.uk

30 May 2011

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing.
Confidence Estimation for Machine Translation.
In *20th Coling*, pages 315–321, Geneva, 2004.

Chih-Chung Chang and Chih-Jen Lin.
*LIBSVM: a library for support vector machines*, 2001.
Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

S. Gandrabur and G. Foster.
Confidence estimation for translation prediction.
In *7th Conference on Natural Language Learning*, pages 95–102, Edmonton, 2003.

Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way.
Bridging smt and tm with translation recommendation.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, Uppsala, Sweden, July 2010.

Y. Kadri and J. Y. Nie.

Improving query translation with confidence estimation for cross language information retrieval.

In *15th ACM International Conference on Information and Knowledge Management*, pages 818–819, Arlington, 2006.

Chris Quirk.

Training a Sentence-Level Machine Translation Confidence Measure.

In *4th Conference on Language Resources and Evaluation*, pages 825–828, Lisbon, 2004.

Radu Soricut and Abdessamad Echihabi.

Trustrank: Inducing trust in automatic translations via ranking.

In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July 2010.

Lucia Specia and Atefeh Farzindar.

Estimating machine translation post-editing effort with hter.

In *Proceedings of the AMTA 2010 Workshop Bringing MT to the User: MT Research and the Translation Industry*, 2010.

Lucia Specia, Dhwaj Raj, and Marco Turchi.
Machine translation evaluation versus quality estimation.
*Machine Translation*, pages 1–12, 2010.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini.
Estimating the Sentence-Level Quality of Machine Translation Systems.
In *13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona, 2009.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders.
Improving the confidence of machine translation quality estimates.
In *Proceedings of the Machine Translation Summit XII*, August 2009.

N. Ueffing and H. Ney.
Application of word-level confidence measures in interactive statistical machine translation.
In *10th Meeting of the European Association for Machine Translation*, pages 262–270, Budapest, 2005.