Project funded by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

Project reference: 317471

Project full title: EXPloiting Empirical appRoaches to Translation

D5.1: Framework for quality estimation

Authors: Carolina Scarton (USFD)

Contributors: Lucia Specia (USFD), Josef van Genabith (USAAR), Marcos Zampieri (USAAR)

Document Number: EXPERT_D5.1_20150510

Distribution Level: Public

Contractual Date of Delivery: 30.10.14

Actual Date of Delivery: 10.05.15

Contributing to the Deliverable: WP5

WP Task Responsible: USFD

EC Project Officer: Concepcion Perez-Camaras

D5.1:Framework for quality estimation WP5: Informing users and learning from user feedback

ESR7: Carolina Scarton
Supervisor: Dr Lucia Specia
Co-supervisor: Professor Josef van Genabith
University of Sheffield
211 Portobello, Sheffield - UK, S1 4DP
c.scarton@sheffield.ac.uk

May 10, 2015

Abstract

One of the main user criticisms of current Machine Translation (MT) technologies is that translators have to repeatedly correct the same errors in translation systems' output over time. The indirect feedback given by correcting translations is not taken into account. MT end-users also criticise that the systems provide no information about the quality of translated segments. In this context, Quality Estimation (QE) task tries to overcome these shortcomings. In the area of machine translations, QE is a kind of evaluation that only considers source and target texts, without the need for a reference translation. This approach is based on using machine learning techniques to predict the quality of unseen data, generalising from a few labelled data points. One research question in QE is trying to find the best features that reduce prediction error. Several features have been used so far, and recent work has tried to include linguistic information. However, traditional QE research presents sentence-level QE evaluation and prediction, disregarding document-level information. In addition, frameworks that encompass feature extraction and prediction have been proposed so far for sentence level. In this report, we present extensions made into QuEst framework. This framework contains two main modules: feature extraction and machine learning. Although it was originally designed for sentence-level QE, other levels are also being addressed and our extensions were made in order to support document-level QE. We also present our document-level features that use lexical cohesion information (a kind of discourse phenomena). This work was done in the context of the **WP5: Informing users and learning from user feedback** of EXPERT project, more specifically, the deliverable **5.1 - Framework for quality estimation** is reported herein.

Contents

1 Introduction

One challenge in Natural Language Processing (NLP) is **how to automatically evaluate language output tasks** such as Machine Translation (MT) and Automatic Summarisation (AS). Although the nature of these tasks is different, they are related in the sense that a "target" text is produced given an input "source" text. Evaluation metrics for these tasks should be able to measure quality with respect to different aspects (e.g. fluency and adequacy) and they should be fast and scalable. Human evaluation seems to be the most reliable (although it might introduce biases of reviewers). However, it is expensive and cumbersome for large datasets; it is also not practical for certain scenarios, such as *gisting* in MT and summarisation of webpages. Therefore, a significant amount of work has targeted measuring quality of MT and AS without direct human effort.

BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin and Och, 2004) are widely used automatic evaluation metrics for MT and AS. These metrics compare the outputs of MT or AS systems with human reference translations or summaries. BLEU is a precision-oriented metric that compares n-grams (typically n=1..4) from reference documents against n-grams in the MT output, measuring how close the output of a system is to one or more references. TER (Translation Error Rate) measures the minimum number of edits required to transform the MT output into the closest reference document. METEOR (Metric for Evaluation of Translation with Explicit ORdering) scores MT outputs by aligning them with given references. This alignment can be done by exact, stem, synonym and paraphrases matching. ROUGE is a recall-oriented metric that measures similarity between sentences by considering the longest common n-gram statistics between a system output sentence and the corresponding reference text.

One limitation of these metrics is that if the MT or AS system outputs a translation or summary considerably different from the references, it does not really mean that it is a bad output. Another problem is that human effort is still needed to produce the references. Finally, and more importantly, these metrics cannot be used in scenarios where the output of the system is to be used directly by end-user, for example a user reading the output of Google Translate[1] for a given news text cannot count on a reference for that translated text. In the context of MT, in this deliverable we focus on approaches for **Quality Estimation (QE)**.

QE approaches aim to predict the quality of MT systems without using references, only features (that may be or may not be related to the MT system that produced this translations) are applied to source and target documents (Blatz et al., 2004; Specia et al., 2009; Bojar et al., 2013). The only requirement is data points with scores (e.g.: Human-targeted Translation Error Rate (HTER) (Snover et al., 2006) or even BLEU-style metrics) to train supervised machine learning models (regressors or classifiers) to predict the scores of unseen data. The advantage of these approaches is that we do not need to have all the words, sentences or documents of a task evaluated manually, we just need enough data points to build the machine learning model. QE systems predict scores that reflect how good a translation is for a given scenario. For example, a widely

[1] https://translate.google.com/

predicted score in QE is HTER, that measures the effort needed to post-edit a sentence.

Most current work on QE is done at the sentence level: the data points are sentences, and a quality label is provided for each of them. A popular application of sentence-level QE is to support post-editing of machine translations (He et al., 2010). As quality labels, Likert scores for post-editing effort, post-editing time, or HTER have been used.

There are, however, scenarios where quality prediction beyond sentence level is needed, most notably in cases when automatic translations without post-editing are required. This is the case, for example, of quality prediction for an entire product review translation in order to decide whether or not it can be published as is, so that customers speaking other languages can understand it. In this report we focus on report our research for document level QE and a framework that encompass word-, sentence- and document-level features extraction and a pipeline to use lower level prediction as features for higher level QE training (such as use word-level prediction for sentence-level QE).

QE for MT has a number of challenges (focusing on document level):

Use of Linguistic Information A challenge in QE is "how do we use linguistic information?" to improve predictions. For sentence-level QE, previous work has explored linguistic information at several levels (such as syntactic and semantic) (Avramidis et al., 2011; Pighin and Màrquez, 2011; Hardmeier, 2011; Felice and Specia, 2012; Almaghout and Specia, 2013). **Discourse** is a linguistic phenomenon that often manifests document-wide. It is related to how sentences are connected, how genre and domain of a document are identified, anaphoric pronouns, etc. Since the state-of-the-art MT systems translate documents at sentence-level, disregarding discourse information, it is expected that the outputs of these systems may contain discourse problems. Because of that, recently there have been initiatives to include discourse information in MT (Marcu et al., 2000; Carpuat, 2009; Zhengxian et al., 2010; LeNagard and Kohen, 2010; Meyer and Popescu-Belis, 2012; Ture et al., 2012; Ben et al., 2013a; Hardmeier, 2014), MT evaluation (Giménez and Màrquez, 2009; Giménez et al., 2010; Wong and Kit, 2012; Meyer et al., 2012; Guzmán et al., 2014) and also in Quality Estimation (Rubino et al., 2013; Scarton and Specia, 2014). Discursive features have also proved to be useful in the evaluation of other Natural Language Processing (NLP) tasks such as Readability Assessment.

Granularity level The vast majority of work done on QE is at **sentence-level** (Specia et al., 2009; Bojar et al., 2014, 2013; Callison-Burch et al., 2012). Going back to the post-editing effort example, sentence-level approaches are very useful in this scenario and in many others (e.g. gisting, mixing of MT systems). Moreover, not only are the predictions made at sentence-level, but also the features are extracted at this level. **Word-level** QE have also been addressed so far (Bojar et al., 2013, 2014). **Document-level** predictions have been only marginally explored (Soricut and Echihabi, 2010; Soricut et al., 2012; Scarton and Specia, 2014). This level of QE is interesting in scenarios where one wants to evaluate the overall score of an MT system or where the end-user is interested in the quality of the document as whole. In addition, document-level features can also correlate well with quality scores, mainly because Statistical

Machine Translation (SMT) translates at sentence-level (it is expected to find several errors beyond sentence boundaries).

Quality labels A challenge of document-level QE is to choose the right quality score to predict. Whilst word-level and sentence-level QE can use human targeted labels (such as likert and HTER) directly, defining guidelines for humans evaluating the document as whole is a very difficult task. One reason is that it is difficult to find the boundaries between sentence-level and document-level problems. Another problem is that evaluation in terms of discourse phenomena (such as cohesion and coherence) is subjective. Up to now, traditional evaluation metrics have been used for document-level prediction. However, these metrics tend to yield similar scores for different documents. This leads to low variation between the document quality scores and all these scores are close to the mean score. Our hypothesis is that traditional metrics, developed to evaluate outputs of different MT systems, do not capture differences of documents translated by a unique system, because they focus on problems that a MT system tends to consistently repeat for all documents. In addition, we claim that a good document-level quality label is not a simple combination of sentence-level quality scores. Scarton et al. (2015) propose a way to assess documents, in order to achieve quality labels for QE. They propose a two-stage post-edition: in the first stage, annotators post-edit sentences randomly presented without context. On the second step, the sentences are put together (in paragraph context) and the annotators are asked to correct problems that could only be solved with context.

Features Several features have been explored for sentence-level QE, including shallow (e.g. number of tokens and language model perplexity) and deep linguistic features (e.g. syntactic trees and semantic roles) and also features that take into account MT system information (e.g. n-best list). For word-level QE, features focus on word alignment, lexical, syntactic and semantic. The state-of-the-art features for document-level QE are based on pseudo-reference (Soricut and Echihabi, 2010; Soricut et al., 2012; Soricut and Narsale, 2012; Shah et al., 2013). **Pseudo-references** are translations produced by MT systems other than the system we want to predict the quality for. They are used as "artificial" references to evaluate the output of the MT system of interest. They have also been used for other purposes, e.g., to fulfil the lack of human references available in reference-based MT evaluation (Albrecht and Hwa, 2008) and automatic summary evaluation (Louis and Nenkova, 2013). The application we are interested in, originally proposed by Soricut and Echihabi (2010), is to generate features for QE. In this scenario, reference-based evaluation metrics, such as BLEU, are computed between the MT system output and the pseudo-references, and used to train quality prediction models. However, pseudo-reference features can not be applied in all scenarios (since they need an extra MT system). Therefore, we focus on implementing baseline features for document-level QE and we also experiment lexical cohesion features.

It is important to notice that tools like Asiya toolkit[2] (Giménez and Màrquez, 2010) and QuEst framework[3] (Specia et al., 2013) are available for QE at sen-

[2]http://nlp.lsi.upc.edu/asiya/
[3]http://www.quest.dcs.shef.ac.uk

tence level. Asiya is a toolkit that extracts automatic metrics in order to evaluate MT quality. Besides the traditional metrics (such as BLEU) this toolkit supports many others metrics (at segment and document levels). Finally, Asiya also extracts **confidence estimation** metrics, that evaluate translation quality and translation difficulty.

On the other hand, QuEst is a framework specifically designed for QE. It has modules to extract several features for QE from source and target documents and to experiment with Machine Learning (ML) techniques for predicting QE. Features are divided in two types: glass-box (dependent on the MT system) and black-box (independent on the MT system). The majority of the features are at sentence level, although a new version of QuEst will encompass word and document levels support. In this report, we present the extensions being made in QuEst in order to support document-level QE feature extraction.

This work is part of a Marie Curie FP7 project called EXPERT[4] (EXPlointing Empirical appRoaches to Translation - ITN No. 317471) which aims to train young and experienced researchers to promote the research, development and use of hybrid language translation technologies. This report is part of the *WP5: Informing users and learning from user feedback*, deliverable **5.1 - Framework for quality estimation**.

This report is organised as follows. Section 2 presents the background behind our research in document-level QE. Section 3 contains a description of the existing frameworks for QE, including the object of this report: QuEst framework. In Section 4, the extension made in QuEst in order to support document-level feature extraction are presented. An experiment and results with these features are presented in Section 5. Final remarks are shown in Section 6.

2 Background

In this section we present the Quality Estimation general framework and related work to the use of linguistic features for QE (Section 2.1). Related work to discourse in MT is also presented (Section 2.2).

2.1 Quality Estimation

Previous work on QE has used supervised ML approaches (mainly regression algorithms). Besides the specific ML method adopted, the choice of features is also a design decision that plays a crucial role.

Figure 1 shows the general framework of QE. Words, sentences or documents from source and target and also information from the MT system are used for designing features. The features extracted are used as input to train a QE model. In this training phase supervised ML techniques, such as regression, can be applied. A training set with quality labels is provided for an ML model. These quality labels are the scores that the QE model will learn to predict. Therefore, the QE model will be able to predict a quality score for a new, unseen data points. The quality labels can be *likert* scores, HTER, BLEU, just to cite some widely used examples. Also the ML algorithm can vary (Support Vector Regression Machines - SVR (Drucker et al., 1997) - and Gaussian Process

[4]http://expert-itn.eu/

- GP (Rasmussen and Williams, 2006) - are the state-of-the-art algorithms for QE at sentence level).

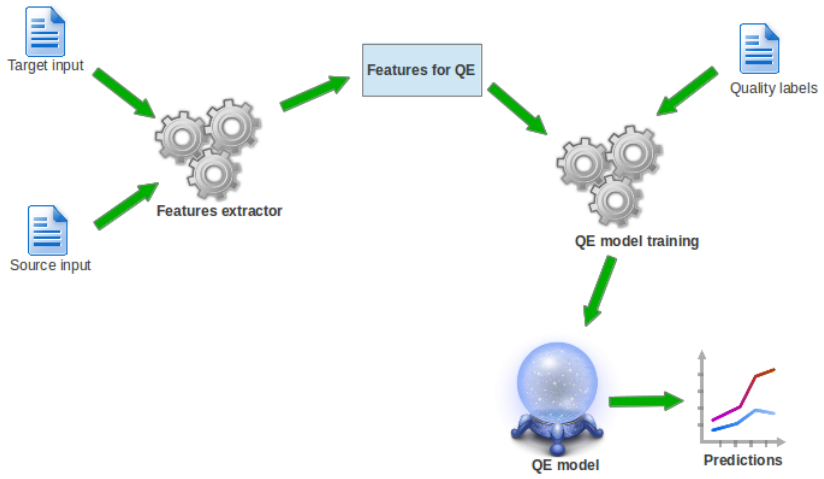

Figure 1: General framework of Quality Estimation

Some work in the area include linguistic information as features for QE (Avramidis et al., 2011; Pighin and Màrquez, 2011; Hardmeier, 2011; Felice and Specia, 2012; Almaghout and Specia, 2013) at sentence level. These features include lexical, syntactic and semantic levels. However, only Scarton and Specia (2014) (predicting quality at document level) and Rubino et al. (2013) (sentence level) focus on the use of discourse information for QE.

At document level, Soricut and Echihabi (2010) explore document-level QE prediction to rank documents translated by a given MT system, predicting BLEU scores. Features include text-based, language model-based, pseudo-reference-based, example-based and training-data-based. Pseudo-reference features are BLEU scores based on pseudo-references from an off-the-shelf MT system, for both the target and the source languages.

Scarton and Specia (2014) explore lexical cohesion and LSA (Latent Semantic Analysis) (Landauer et al., 1998) cohesion for document-level QE. The lexical cohesion features are repetitions (Wong and Kit, 2012) and the LSA cohesion is achieved following the work of Graesser et al. (2004). Pseudo-reference features are also applied in this work, according to the work of Soricut and Echihabi (2010). BLEU and TER (Snover et al., 2006) are used as quality labels. The best results were achieved with pseudo-reference features. However, LSA cohesion features alone also showed improvements over the baseline.

2.2 Discourse use in MT

In the MT area, there have been attempts to use discourse information that can be used as inspiration source for QE features. The need of document-level information for improving MT is a widely accepted fact. However, it is hard to integrate discourse information into traditional state-of-the-art sentence-level

MT systems. It is also challenging to build a document-level or discourse-based MT system from scratch. Therefore, the initiatives focus on the integration of discourse as features into the decoding phase or previously annotate discourse phenomena in the parallel corpora.

Lexical Cohesion is related to word usage: word repetitions, synonyms repetitions and collocations. Besides initiatives to improve MT system and outputs with lexical cohesion (Ture et al., 2012; Xiao et al., 2011; Ben et al., 2013b), Wong and Kit (2012) apply lexical cohesion metrics for evaluation of MT systems at document level.

Coreference is related to coherence clues, such as pronominal anaphora and connectives. Machine translation can break coreference chains since it is done at sentence level. Initiatives for improvement of coreference in MT include anaphora resolution (Giménez et al., 2010; LeNagard and Kohen, 2010; Hardmeier and Federico, 2010; Hardmeier, 2014) and connectives (Popescu-Belis et al., 2012; Meyer and Popescu-Belis, 2012; Meyer et al., 2012; Li et al., 2014).

RST (Rhetorical Structure Theory) (Mann and Thompson, 1987) is a linguistic theory that correlates macro and micro units of discourse in a coherent way. The correlation is made among EDUs (Elementary Discourse Units). EDUs are defined at sentence, phrase or paragraph-level. These correlations are represented in the form of a tree. Marcu et al. (2000) explore RST focusing on identifying the feasibility of building a discourse-based MT system. Guzmán et al. (2014) use RST trees comparison for MT evaluation.

Topic models capture word usage, although they are more robust than lexical cohesion structures because they can correlate words that are not repetitions or do not present any semantic relation. These methods can measure if a document follows a topic, is related to a genre or belongs to a specific domain. Work on improving MT that uses topic models include Zhengxian et al. (2010) and Eidelman et al. (2012).

3 QE frameworks

As mentioned in Section 1, frameworks were developed in order to put together MT evaluation and quality estimation features. **Asiya** is a complete toolkit that encompass many automatic evaluation metrics. These metrics go from shallow levels (such as basic n-gram matches) to deep linguistic information (syntactic trees, semantic and discourse relations, etc). In terms of QE, Asiya have features that evaluate translation quality and translation difficult.

Translation quality is assessed via target-based features (language model perplexity and log probability and out-of-vocabulary tokens ratio) and source/target-based features (bilingual dictionary overlap, ratio between source and target sentence length, linguistic elements overlap and symbol matches).

On the other hand, translation difficulty is evaluated by source-based features (bilingual dictionary ambiguity, language model perplexity and log probability, source length and out-of-vocabulary tokens ratio). In this scenario, the source is assessed in order to define "how difficult" it should be to translate a source word or sentence.

While Asiya is a toolkit mainly designed for MT automatic metrics extraction, with a module for QE metrics, **QuEst** is a framework completely designed for QE purposes. Written in Java, it was originally designed for sentence -level

QE, although it is being extended to other granularity levels. It is basically divided into two main modules: feature extraction and machine learning. Features can be extracted from source and target texts and they can assess complexity, fluency, adequacy and confidence.[5]

Sentence level QuEst extracts glass-box (MT system related) and black-box (MT system independent) features at sentence level. Examples of these features are:

- number of tokens in source (target) sentence;
- language model (LM) probability of source segment using the source side of the parallel corpus used to train the MT system as LM;
- LM probability of target segment using a large corpus of the target language to build the LM;
- ratio of number of tokens in source and target segments;
- difference between the depth of the syntactic trees of the source and target segments;
- difference between the number of person/location/organization entities in source and target sentences;
- features and global score of the SMT system (glass-box);
- 1–3-gram LM probabilities using translations in the n-best to train the LM (glass-box).

The machine learning module was developed in python and allows users to use the extracted features to training QE models. A selection of 17 features from QuEst have been used as baseline in several previous work including the WMT sentence-level QE shared tasks from 2012 to 2015.

A simplified version of QuEst architecture for sentence-level feature extraction and prediction is presented in Figure 2. This figure shows a class called *Sentence* as the basic class in the framework. This class contains information about the sentence (such as number of words), the raw sentence and the PoS tags for each token. For each sentence in the source or target document, an object of *Sentence* is created. Then, each object is processed by a *FeatureExtractionModule* and features are extracted according to the preferences of the user (given via a configuration file). If desirable, the ML module can be used to train a QE model with the extracted features.

Word level Although QuEst has been designed for sentence-level QE, recent efforts have extended QuEst to other levels. Word-level QE is being implemented in QuEst by using the same structure for sentence level.[6] The idea is that, for each sentence, QuEst processes the words and returns a set of features for each word. For word-level prediction, a Conditional Random Fields (CRF) (Lafferty et al., 2001) algorithm is being integrated into the ML module. Examples of features at word level include:

[5]It is worth mentioning that QuEst is now available as a plug-in of OKAPI framework (Paetzold et al., 2015).

[6]https://github.com/ghpaetzold/quest

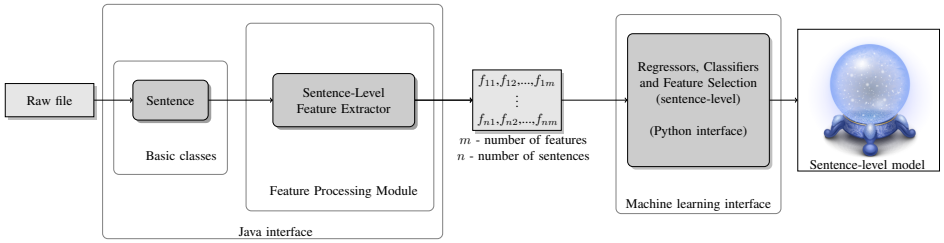

Figure 2: Architecture of QuEst (sentence level).

- Target context: for a given word t_i in a given target sentence, features include the word itself, bigrams and trigrams that contains;
- Alignment context: for a given word t_i in target aligned to a word s_j , features include the aligned word s_j, target-source bigrams and source-target bigrams;
- Lexical: features are POS tags of target words and POS tags of bigrams and trigrams containing these words;
- LM: features are related to the backoff behaviour of a word's context with respect to an LM (Raybaud et al., 2011). Two features are extracted: lexical backoff behaviour and syntactic backoff behaviour;
- Syntactic: Null Link - a binary feature that receives value 1 if a given word t_i in a target sentence has at least one dependency link with another word t_j, and 0 otherwise;
- Semantic: features explore the polysemy of target and source words, i.e. the number of senses existing as entries in a WordNet for a given word t_i or a source word s_i;
- Pseudo-reference: feature explores the similarity between the target sentence and a translation for the source sentence produced by another MT system.

It is worth mentioning that word-level QE is being used by ESR6 in order to develop a *Framework for learning from human translators* (deliverable 5.2) Logacheva and Specia (2015).

In the next section we show the contribution of ESR7 in the development of QuEst framework by including document-level into its architecture.

4 QuEst: extensions at document level

The extension of QuEst at document level was beyond the development of features. The architecture also needed to be changed in order to support the new level. In this section we describe the features implemented so far and the new architecture proposed.

4.1 Features

Document-level features implemented in QuEst are the adaptation of the 17 baseline features at sentence level plus nine lexical cohesion features. The set of 17 document-level baseline features (hereafter, DL-QuEst) is:

- number of tokens in the source document
- number of tokens in the target document
- average source token length
- LM probability of source document (average of sentence-level LM probabilities)
- LM probability of target document (average of sentence-level LM probabilities)
- number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio)
- average number of translations per source word in the sentence (as given by IBM 1 table thresholded such that prob(t—s) ¿ 0.2)
- average number of translations per source word in the sentence (as given by IBM 1 table thresholded such that prob(t—s) ¿ 0.01) weighted by the inverse frequency of each word in the source corpus
- percentage of unigrams in quartile 1 of frequency (lower frequency words) in a corpus of the source language (SMT training corpus)
- percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language
- percentage of bigrams in quartile 1 of frequency of source words in a corpus of the source language
- percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language
- percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language
- percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language
- percentage of unigrams in the source document seen in a corpus (SMT training corpus)
- number of punctuation marks in the source document
- number of punctuation marks in the target document

Lexical cohesion is a discourse phenomenon related to word repetitions and collocation. This phenomenon was explored as features for Readability Assessment (Graesser et al., 2004) and MT evaluation (Wong and Kit, 2012). Following these work, we proposed the first set of features for QE using lexical cohesion (hereafter, LC).

These features are based in words repetitions only. The reason for that is the aim of keep QE as language independent as possible. Synonyms and other kind of semantic relations require the need of resources like WordNet that are not freely available for several languages. Besides that, the coverage of these kind of resources vary across languages, and it could influence in the liability of the feature.

- content words repetition in source and target documents
- lemmas repetition in source and target documents
- nouns repetition in source and target documents
- ratio of content words/lemmas/nouns in source and target documents (three features)

4.2 Architecture

In order to implement document-level feature extraction in QuEst, its architecture needed to be adapted. Whilst sentence- and word-level QE in QuEst use *Sentence* class as their basic class, document-level QE needs to rely on a *Document* class. One of our aims was also to use the already implemented functionalities of QuEst, such as features at sentence level that could be extended to document level. Therefore, the *Document* class contains a list of *paragraphs*, each paragraph being an object of *Paragraph* class. The *Paragraph* class encompass a list of sentences, each sentence being an object of *Sentence* class. In this scenario, a *document* is a set of *paragraphs* that is a set of *sentences*.

This architecture is flexible in the sense that a document can be considered a combination of paragraphs or sentences (via class objects) or not (by concatenating all sentences together). It is also robust to paragraph-level QE (one could only implement features for this level) and paragraph-driven features for document-level (macro unit features, such as defining the purposes of a given paragraph: introduction, background, conclusion, etc).

Figure 3 shows the architecture for document-level feature extraction and prediction. As mentioned previously, additional classes for document and paragraphs were created. Different from word-level QE, document-level QE can use the same ML pipeline as sentence-level.

Another change that needs to be made in QuEst is how it deals with input files. For word- and sentence-level feature extraction, QuEst receives a raw file, with a sentence per line. However, since we need to deal with documents, a more robust file structure should be supported. We, then, propose the use of SGML files as input for document-level feature extraction. This kind of file can contain several documents with paragraphs and sentences mark-ups. The choice of this format is also supported by the fact that WMT shared tasks use it.

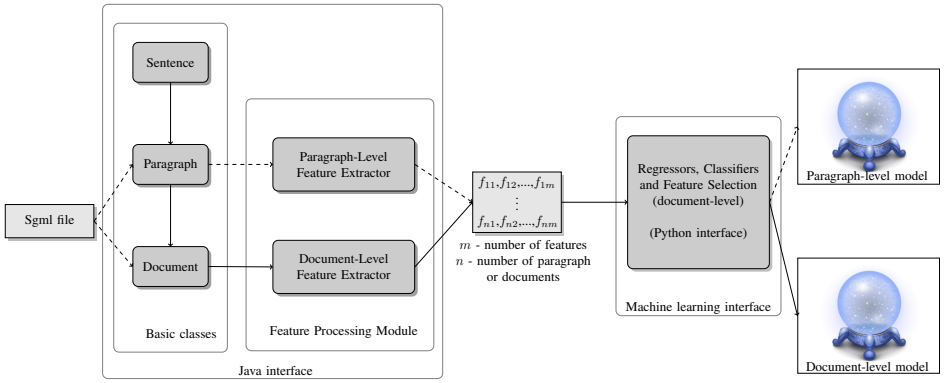

Figure 3: Architecture of QuEst with document-level support. Dashed lines are for work in progress

5 Experiments and Results

We conducted several experiments with baseline and LC features at document level. In this section we show results of our experiments that also included two other sets of features:

LSA features Latent Semantic Analysis (**LSA**) Landauer et al. (1998) is used in order to extract cohesion-related features. This is a statistical method based on Singular Vector Decomposition (SVD) and is often aimed at dimensionality reduction. The LSA matrix (rows x columns) can be built from words by sentences, words by documents, sentences by documents, etc. In the case of words by sentences (which we use in our experiments), each cell contains the frequency of a given word in a given sentence. LSA was originally designed to be used with large corpora of multiple documents. In our case, since we are interested in measuring cohesion within documents, we compute LSA for each individual document through a matrix of words by sentences within the document. Our LSA features follow from Graesser et al. (2004)'s work on readability assessment:

- LSA adjacent sentences: for each sentence in a document, we compute the Spearman rank correlation coefficient of its word vector with the word vectors of its immediate neighbours (sentences which appear immediately before and after the given sentence). For sentences with two neighbours (most cases), we average the correlation values. After that, we average the values for all sentences in order to have a single figure for the entire document.
- LSA all sentences: for each sentence in a document, we calculate the Spearman rank correlation coefficient of the word vectors between this sentence and all the others. Again we average the values for all sentences in the document.

Higher correlation scores are expected to correspond to higher text cohesion, since the correlation among the sentences in a document is related to how close the words in the document are Graesser et al. (2004). Different from lexical cohesion features, LSA features are able to find correlations among different words, which are not repetitions and may not be synonyms, but are instead related (as given by co-occurrence patterns).

Pseudo-reference Pseudo-references are translations produced by other MT systems than the system we want to predict the quality for. They are used as references to evaluate the output of the MT system of interest. The application we are interested in, originally proposed in Soricut and Echihabi (2010), is to generate features for QE. In this scenario, reference-based evaluation metrics (such as BLEU) are computed between the MT system output and the pseudo-references and used to train quality prediction models. In our experiments, BLEU and TER scores are computed between the output of the MT system of interest and alternative MT systems, at document-level, and used as features in QE models.

5.1 Experimental Settings

Although QE is traditionally trained on datasets with human labels for quality (such as HTER – Human Translation Error Rate Snover et al. (2006)), no large enough dataset with human-based quality labels assigned at document-level is available. Therefore, we resort to predicting automatic metrics as quality labels, as in Soricut and Echihabi (2010). This requires references (human) translations at training time, when the automatic metrics are computed, but not at test time, when the automatic metrics are predicted.

Corpora Two parallel corpora with reference translations are used in our experiments: FAPESP and WMT13. **FAPESP** contains 2,823 English-Brazilian Portuguese (EN-BP) documents extracted from a scientific Brazilian news journal (FAPESP)[7] Aziz and Specia (2011). Each article covers one particular scientific news topic. The corpus was randomly divided into 60% (1,694 documents) for training a baseline **MOSES**[8] statistical MT system Koehn et al. (2007) (with 20 documents as development set); and 40% (1,128 documents) for testing the SMT system, which generated translations for QE training (60%: 677 documents) and test (40%: 451 documents). In addition, two external MT systems were used to translate the test set: **SYSTRAN**[9] – a rule-based system – and Google Translate (**GOOGLE**), a statistical system.

WMT13 contains English-Spanish (**EN-ES**) and Spanish-English (**ES-EN**) translations from the test set of the translation shared task of WMT13.[10] In total, 52 source documents were available for each language pair. In order to build the QE systems, the outputs of all MT systems submitted to the shared task were taken: 18 systems for EN-ES (528 documents for QE training, and 356 for QE test), and 17 systems for ES-EN (500 documents for QE training,

[7] `http://revistapesquisa.fapesp.br`
[8] `http://www.statmt.org/moses/?n=moses.baseline`
[9] `http://www.systransoft.com/`
[10] `http://www.statmt.org/wmt13/`

and 332 documents for QE test). In both cases, the translations from one MT system are used as pseudo-references for translations from the other systems.

Quality labels The automatic metrics selected for quality labelling and prediction are BLEU and TER.[11] **BLEU** (BiLingual Evaluation Understudy) is a precision-oriented metric that compares n-grams (n=1-4 in our case) from reference documents against n-grams of the MT output, measuring how close the output of the system is to one or more references. **TER** (Translation Error Rate) Snover et al. (2006) measures the minimum number of edits required to transform the MT output in the reference document. The Asiya Toolkit[12]Giménez and Màrquez (2010) was used to calculate both metrics.

Baseline As a baseline (**Mean**), we calculate the average BLEU or TER scores in the QE training set, and apply this value to all entries (documents) in the test set.

Pseudo-reference features For the FAPESP corpus, translations from Google Translate were selected as pseudo-references, since this system has shown the best average BLEU score in the QE training set. For the WMT13 corpus, translations from *uedin-wmt13-en-es*, for EN-ES, and *uedin-heafield-unconstrained* for ES-EN, were used as pseudo-references, since these systems achieved the best BLEU scores in the WMT13 translation shared task. Regarding the difference between the systems, for the FAPESP corpus, this difference is guaranteed since GOOGLE is considerably different from SYSTRAN, and is trained on a different (much larger) corpus than MOSES. For the WMT13 corpus, it is not possible to make this assumption, as many of the systems participating in the shared task are close variations of Moses.

Feature sets As feature sets, we combine LC and LSA features with DL-QuEst (**DL-QuEst+LC**, **DL-QuEst+LSA** and **DL-QuEst+LC+LSA**) to create the models with discursive information. The pseudo-reference features are combined with the DL-QuEst (**DL-QuEst+Pseudo**) and with all other features (**DL-QuEst+LC+LSA+Pseudo**).

Machine learning algorithm We use the SVM regression algorithm (SVR) with a radial basis function kernel and hyperparameters optimised via grid search to train the QE models with all feature sets The scikit-learn module available in QuEst was used for that.

Evaluation metrics The QE models with different feature sets are evaluated using **MAE** (Mean Absolute Error): $MAE = \frac{\sum_{i=1}^{n} |H(s_i) - V(s_i)|}{N}$ where $H(s_i)$ is the predicted score, $V(s_i)$ is the true score and N is the number of data points in the test set. To verify the significance of the results, two-tailed pairwise t-test ($p<0.05$) was performed for different prediction outputs.

[11] METEOR was also used but the results were inconclusive

[12] `http://asiya.lsi.upc.edu/`

Method Two sets of experiments were conduct. First (Section 5.2.1), we consider the outputs of the FAPESP corpus of MOSES, SYSTRAN and GOOGLE separately, using as training and test sets the outputs of each system individually, with GOOGLE translations used as pseudo-references for the other two systems. The second set of experiments (Section 5.2.2) considers, for the FAPESP corpus, the combination of the outputs of MOSES and SYSTRAN (MOS+SYS), again with GOOGLE translations used as pseudo-references. For the WMT2013 corpora, we mixed translations from all except the best system, which were used as pseudo-references.

5.2 Results

5.2.1 MT system-specific models

The results for the prediction of BLEU and TER for MOSES, SYSTRAN and GOOGLE systems in the FAPESP corpus are shown in Table 1. The best results for MOSES and SYSTRAN were obtained with the inclusion of pseudo-references (DL-QuEst+Pseudo and DL-QuEst+LC+LSA+Pseudo), with both BLEU and TER. However, only the improvements for MOSES showed statistically significant difference: with both BLEU and TER, the best results were tied between DL-QuEst+Pseudo and DL-QuEst+LC+LSA+Pseudo, but there are still significant differences between their predictions. An interesting finding is that without considering pseudo-reference features for MOSES and SYSTRAN, the best results are achieved with LSA features. In fact, for SYSTRAN the results from using of only DL-QuEst+LSA are not significantly different from the use of all features (including pseudo-references).

For GOOGLE, the best results (for BLEU and TER) were obtained by DL-QuEst+LC [13]. However, BLEU predictions showed no significant difference among all feature sets and the best TER figure was not significantly different from DL-QuEst+LC+LSA.

	BLEU			TER		
	MOSES	SYSTRAN	GOOGLE	MOSES	SYSTRAN	GOOGLE
Mean	0.059	0.047	0.066	0.063	0.062	0.068
DL-QuEst	0.046	0.047	0.056	0.054	0.059	0.061
DL-QuEst+LC	0.044	0.043	0.055	0.053	0.059	0.055
DL-QuEst+LSA	0.044	0.044	0.058	0.055	0.059	0.060
DL-QuEst+LC+LSA	0.044	0.043	0.057	0.053	0.058	0.061
DL-QuEst+Pseudo	0.042*	0.038	-	0.052*	0.051	-
DL-QuEst+LC+LSA+Pseudo	0.042*	0.036	-	0.052*	0.051	-
Test-set average	0.365	0.275	0.456	0.427	0.506	0.372
Test-set range	[0.004,0.558]	[0,0.406]	[0.004, 0.79]	[0.245,1.056]	[0,1.071]	[0.12,1.084]

Table 1: MAE scores for document-level prediction of BLEU and TER for the FAPESP corpus. Bold-faced figures indicate the smallest MAE for a given test set; * indicates a statistically significant difference against all other results; underlined values indicate no significant difference against the best system.

In order to understand whether the MAE scores obtained are "good enough", it is interesting to compare them against the error of the Mean baseline, but also to analyse the average of the true scores and the range of variation of these true scores in the test set (last two lines in Table 1). For the prediction of BLEU

[13] Pseudo-reference features were not used for GOOGLE, since its outputs was used as pseudo-reference for the other systems.

scores, the true scores range from 0 to 0.5 for MOSES and SYSTRAN, and from 0 to 0.8 for GOOGLE. This suggests that the impact of error differences in MOSES and SYSTRAN is higher. A wider range of scores and a relatively higher Mean MAE could indicate a relatively easier prediction task. This is directly connected to the variation in the quality of the translations in the datasets. This seems to be the case with BLEU prediction for GOOGLE translations: the improvements between the Mean baseline and the DL-QuEst features is much higher than with the other MT systems. The variation in terms of TER is larger, making improvements over the Mean baseline possible with all feature sets.

Given the low MAE scores obtained by the Mean baseline, as well as with simple DL-QuEst features, one could say that in general the task of predicting BLEU and TER is close to trivial, at least in the FAPESP corpus. This is again due to the low variation in the quality of texts translated by each system. This is to be expected, given the very nature of document-level prediction: major variations in the quality of specific translated segments get smoothed out throughout the document. In addition, the FAPESP corpus consists of texts from the same style and domain. On the other hand, the average quality (as measured by BLEU and TER metrics) of the different MT systems on the same corpus is very different, as shown in the penultimate line of Table 1. This motivates the experiment described next.

5.2.2 MT system-independent models

To analyse document-level QE in a more challenging scenario, we experiment with mixing different MT system outputs, for both FAPESP and WMT2013 corpora. Results are shown in Table 2.

	BLEU			TER		
	FAPESP	WMT2013		FAPESP	WMT2013	
	MOS+SYS	EN-ES	ES-EN	MOS+SYS	EN-ES	ES-EN
Mean	0.064	0.061	0.076	0.07	0.066	0.089
DL-QuEst	0.045	0.056	0.065	0.063	0.059	0.069
DL-QuEst+LC	0.044	0.058	0.065	0.063	0.066	0.07
DL-QuEst+LSA	0.044	0.052	0.051	0.062	0.057	0.051
DL-QuEst+LC+LSA	0.044	0.053	0.052	0.064	0.054	0.062
DL-QuEst+Pseudo	0.043	0.043	0.038	0.053	0.034	0.038*
DL-QuEst+LC+LSA+Pseudo	0.038*	0.045	0.043	0.054	0.034	0.04
Test-set average	0.32	0.266	0.261	0.466	0.524	0.55
Test-set range	[0,0.558]	[0.107,0.488]	[0.072,0.635]	[0,1.07]	[0.317,0.72]	[0.216,0.907]

Table 2: MAE scores for document-level prediction of BLEU and TER for the FAPESP corpus (mixing MOSES and SYSTRAN) and for the WMT2013 EN-ES and ES-EN corpora (mixing all but best system).

The ranges of BLEU/TER scores are now wider, and the overall error scores (including for the Mean baseline) are higher in these settings, showing that this is indeed a harder task. Again, the best results are obtained with the use of pseudo-reference features. However, in this case statistically significant differences against other results were only observed with MOS+SYS BLEU prediction and ES-EN TER prediction. For EN-ES BLEU prediction, the best result (0.043 for DL-QuEst+Pseudo) showed no significant difference against DL-QuEst+LC+LSA+Pseudo (0.045). For ES-EN BLEU prediction, there is no significant difference among the results of DL-QuEst+LSA, DL-QuEst+LC+LSA

and DL-QuEst+Pseudo. For MOS+SYS TER prediction, DL-QuEst+Pseudo and DL-QuEst+LC+LSA+Pseudo showed no significant difference. EN-ES TER prediction was the only case were the DL-QuEst results showed no significant difference against pseudo-reference features. It is worth mentioning that, as in the previous experiments, if we disregard the pseudo-reference features – which may not be available in many real-world scenarios – the LSA feature sets show the best results.

6 Final Remarks

In this report, we describe the deliverable 5.1 of EXPERT project: Framework for quality estimation. We present QuEst, a general framework for QE with support for word-, sentence- and document-level QE feature extraction and QE prediction.

It is worth mentioning that we plan to continue the work with document-level QE by exploring more discourse-related features. We plan to include output of discourse parsers and taggers as features for QE. With the new architecture, even relations beyond sentences (such as paragraphs) can be implemented.

Finally, with the new version of QuEst one can use it to train word-level QE models and use it results as features for sentence- or document-level models (the same can be done for sentence-level QE prediction as features for document-level QE). This will be included as a suggestion for the users of QuEst framework.

References

Albrecht, J. S. and Hwa, R. (2008). The role of pseudo references in mt evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 187–190, Columbus, Ohio, USA.

Almaghout, H. and Specia, L. (2013). A ccg-based quality estimation metric for statistical machine translation. In *Proceedings of the MT Summit 2013*.

Avramidis, E., Popovic, M., Torres, D. V., and Burchardt, A. (2011). Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 65–70, Edinburgh, UK.

Aziz, W. and Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*, Cuiabá, MT.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Ben, G., Xiong, D., Teng, Z., Lü, Y., and Liu, Q. (2013a). Bilingual lexical cohesion trigger model for document-level machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 382–386, Sofia, Bulgaria. Association for Computational Linguistics.

Ben, G., Xiong, D., Teng, Z., Lu, Y., and Liu, Q. (2013b). Bilingual lexical cohesion trigger model for document-level machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 382–386, Sofia, Bulgaria.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *The 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *The Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *The Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *The Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Carpuat, M. (2009). One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, CO.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. In Mozer, M. C., Jordan, J. I., and Petscbe, T., editors, *Neural Information Processing Systems 9*, pages 155–161. MIT Press, Cambridge, Massachusetts.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic models of dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 115–119, Jeju Island, Korea.

Felice, M. and Specia, L. (2012). Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103.

Giménez, J. and Màrquez, L. (2009). On the robustness of syntactic and semantic features for automatic mt evaluation. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, pages 250–258, Athens, Greece.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Giménez, J., Màrquez, L., Comelles, E., Catellón, I., and Arranz, V. (2010). Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36:193–202.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using Discourse Structure Improves Machine Translation Evaluation. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, MD.

Hardmeier, C. (2011). Improving machine translation quality prediction with syntatic tree kernels. In *Proceedings of the 15th conference of the European Assoaciation for Machine Translation (EAMT 2011)*, pages 233–240, Leuven, Belgium.

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis, Department of Linguistics and Philology, Uppsala University, Sweden.

Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 283–289.

He, Y., Ma, Y., van Genabith, J., and Way, A. (2010). Bridging SMT and TM with Translation Recommendation. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, Uppsala, Sweden.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, Marcello, B. N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstration session*, Prague, Czech Republic.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

LeNagard, R. and Kohen, P. (2010). Aiding pronoun translation with coreference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppasala, Sweden.

Li, J. J., Carpuat, M., and Nenkova, A. (2014). Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, pages 283–288, Baltimore, MD.

Lin, C.-Y. and Och, F. J. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of ACL 2004*, Barcelona, Spain.

Logacheva, V. and Specia, L. (2015). The role of artificially generated negative data for quality estimation of machine translation. In *To appear in the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey.

Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Cambridge University Press, Cambridge, UK.

Marcu, D., Carlson, L., and Watanabe, M. (2000). The automatic translation of discourse structures. In *NAACL 2000 Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17. Association for Computational Linguistics.

Meyer, T. and Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France.

Meyer, T., Popescu-Belis, A., Hajlaoui, N., and Gesmundo, A. (2012). Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012)*, San Diego, CA.

Paetzold, G. H., Specia, L., and Savourel, Y. (2015). Okapi+QuEst: Translation Quality Estimation within Okapi. In *To appear in the 18th Annual Conference of the European Association for Machine Translation - demonstration session*, Antalya, Turkey.

Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *The 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Pighin, D. and Màrquez, L. (2011). Automatic projection of semantic structures: an application to pairwise translation ranking. In *Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, number June, pages 1–9.

Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B., and Zufferey, S. (2012). Discourse-level annotation over europarl for machine translation: connectives and pronouns. In *Proceedings of the Eighth Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

Rubino, R., Souza, J. G. C. d., Foster, J., and Specia, L. (2013). Topic models for translation quality estimation for gisting purposes. In *Proceedings of the XIV Machine Translation Summit*, pages 295–302, Nice, France.

Scarton, C. and Specia, L. (2014). Document-level translation quality estimation: exploring discourse and pseudo-references. In *The 17th Annual Conference of the European Association for Machine Translation*, pages 101–108, Dubrovnik, Croatia.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015). Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *To appear in the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey.

Shah, K., Cohn, T., and Specia, L. (2013). An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of the XIV MT Summit*, pages 167–174, Nice, France.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Seventh biennial conference of the Association for Machine Translation in the Americas*, AMTA 2006, pages 223–231, Cambridge, MA.

Soricut, R., Bach, N., and Wang, Z. (2012). The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In *Proceedings of WMT 2012*, pages 145–151, Montréal, Canada.

Soricut, R. and Echihabi, A. (2010). TrustRank: Inducing Trust in Automatic Translations via Ranking. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden.

Soricut, R. and Narsale, S. (2012). Combining Quality Prediction and System Selection for Improved Automatic Translation Output. In *Proceedings of WMT 2012*, pages 163–170, Montréal, Canada.

Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (2013). Quest - a translation quality estimation framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *The 13th Annual Conference of the European Association for Machine Translation*, pages 28–37, Barcelona, Spain.

Ture, F., Oard, D. W., and Resnik, P. (2012). Encouraging consistent translation choices. In *Proceedings of the 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2012)*, pages 417–426, Montreal, Canada.

Wong, B. T. M. and Kit, C. (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068. Association for Computational Linguistics.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 131–138, Xiamen, China.

Zhengxian, G., Yu, Z., and Guodong, Z. (2010). Statistical machine translation based on lda. In *Proceedings of the 4th International Universal Communication Symposium (IUCS 2010)*, pages 279–283, Beijing, China.