Relevance Ranking for Translated Texts

Marco Turchi, Josef Steinberger
European Commission JRC
IPSC - GlobeSec
Via Fermi 2749,
21020 Ispra (VA), Italy
name.surname@jrc.ec.europa.eu

Lucia Specia
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP, UK
L.Specia@dcs.shef.ac.uk

Abstract

The usefulness of a translated text for gisting purposes strongly depends on the overall translation quality of the text, but especially on the translation quality of the most informative portions of the text. In this paper we address the problems of ranking translated sentences within a document and ranking translated documents within a set of documents on the same topic according to their informativeness and translation quality. An approach combining quality estimation and sentence ranking methods is used. Experiments with French-English translation using four sets of news commentary documents show promising results for both sentence and document ranking. We believe that this approach can be useful in several practical scenarios where translation is aimed at gisting, such as multilingual media monitoring and news analysis applications.

1 Introduction

Reading and understanding the main ideas behind documents written in different languages can be necessary or desirable in a number of scenarios. Existing online translation systems such as *Google Translate* and *Bing Translator*[1] serve to this purpose, mitigating the language barrier effects. Despite the large improvements in translation quality in recent years, translated documents are still affected by the presence of sentences which are not correctly translated and in the extreme case, whose original meaning has been lost. These sentences can compromise the readability and reliability of translated documents, especially if they are the ones that should convey the most important information in the document.

Quality estimation methods can flag incorrect translations without access to reference sentences, however the informativeness of these sentences is not taken into account. On the other hand, sentence ranking methods are able to identify the most relevant sentences in a given language for tasks such as document summarisation. However, the performance of sentence ranking algorithms for machine translated texts can be significantly degraded due to the introduction of errors by the translation process, as it has been shown for other language processing tasks, e.g. in information retrieval (Savoy and Dolamic, 2009). Moreover, particularly in the case of supervised ranking methods, these may only be available for the source language.

In this paper we propose combining quality estimation and relevance sentence ranking methods in order to identify the most relevant translated texts. We experiment with two ranking tasks:

- The ranking of translated sentences within a document; and
- The ranking of documents within a set of documents on the same topic.

An evaluation with French-English translations in groups of news commentary documents in different domains has shown promising results for both sentence and document ranking.

2 Related work

A considerable amount of work has been dedicated in recent years to estimating the quality of ma-

[1] translate.google.com/ and www.microsofttranslator.com/

chine translated texts, i.e., the problem of predicting the quality of translated text without access to reference translations. Most related work focus on predicting different types of sentence-level quality scores, including automatic and semi-automatic MT evaluation metrics such as TER (He et al., 2010), HTER (Specia and Farzindar, 2010; Bach et al., 2011), post-editing effort scores and post-editing time (Specia, 2011). At document level, similar to this paper, Soricut and Echihabi (2010) focus on the ranking translated documents according to their estimated quality so that the top *n* documents can be selected for publishing. A range of indicators from the MT system, source and translation texts have been used in previous work. However, none of these include the notion of informativeness of the texts.

The sentence ranking problem has been widely studied in particular for document summarization, where different approaches have been proposed to quantify the amount of information contained in each sentence. In (Goldstein et al., 1999), a technique called Maximal Marginal Relevance (MMR) was introduced to measure the relevance of each sentence in a document according to a user provided query. Other approaches represent a document as a set of trees and take the position of a sentence in a tree is indicative of its importance (Carlson et al., 2001). Graph theory has been extensively used to rank sentences (Yeh et al., 2008) or keywords (Mihalcea, 2004), with their importance determined using graph connectivity measures such as in-degree or PageRank. A sentence extraction method based on Singular Value Decomposition over term-by-sentence matrices was introduced in (Gong and Xin, 2002).

The combination of relevance and translation quality scores has been recently proposed in the context of cross-language document summarization. In (Wan et al., 2010), sentences in a document were ranked using the product of quality estimation and relevance scores, both computed using the source text only. The best five sentences were added to a summary, and then translated to the target language. (Boudin et al., 2010) used both source and target language features for quality estimation and targeted multi-document summarization, selecting sentences from different translated documents to generate a summary.

This paper extends previous work in the attempt to rank translated sentences within documents, but with a different objective: instead of selecting a pre-defined number of sentences to compose a summary, we aim at obtaining a global ranking of sentences within a document according to their informativeness and translation quality and use this ranking to assign a global score to each document for the ranking of groups of documents. This requires different evaluation strategies from those used in the text summarization field, as we will discuss in Section 5.2.

3 Quality estimation method

The quality estimation method used in this paper is that proposed in (Specia, 2011). A sentence-level model is built using a Support Vector Machines regression algorithm with radial basis function kernel from the LIBSVM package (Chang and Lin, 2011) and a number of shallow and MT system-independent features. These features are extracted from the source sentences and their corresponding translations, and from monolingual and parallel corpora. They include source & translation sentence lengths, source & translation sentence language model probabilities, average number of translations per source word, as given by probabilistic dictionaries, percentages of numbers, content-/non-content words in the source & translation sentences, among others. The regression algorithm is trained on examples of translations and their respective human judgments for translation quality (Section 5.1).

4 Sentence ranking methods

4.1 Co-occurrence-based ranking

Originally proposed by (Gong and Xin, 2002) and later improved by (Steinberger and Ježek, 2004), this is an unsupervised method based on the application of Singular Value Decomposition (SVD) to individual documents or sets of documents on the same topic. It has been reported to have the best performance in the multilingual multi-document summarization task at TAC 2011. The method first builds a term-by-sentence matrix from the text, then applies SVD and uses the resulting matrices to identify and extract the most salient sentences. SVD is aimed at finding the latent (orthogonal) dimensions, which would correspond to the different topics discussed in the set of documents.

More formally, we first build a matrix **A** where each column represents the weighted term-frequency vector of a sentence j in a given docu-

ment or set of documents. The weighting schemes found to work best in (Steinberger and Ježek, 2009) are a binary local weight and an entropy-based global weight.

After that step, SVD is applied to the matrix as $\mathbf{A} = \mathbf{USV}^T$, and subsequently a matrix $\mathbf{F} = \mathbf{S} \cdot \mathbf{V}^T$ reduced to r dimensions[2] is derived.

Sentence selection starts with measuring the length of the sentence vectors in $\mathbf{F}$. This length can be viewed as a measure of the importance of that sentence within the top topics (the most important dimensions). In other words, the length corresponds to the combined weight across the most important topics. We call it *co-occurrence sentence score*. The sentence with the largest score is selected as the most informative (its corresponding vector in $\mathbf{F}$ is denoted by $\mathbf{f}_{best}$). To prevent selecting a sentence with similar content in the next step, the topic/sentence distribution in matrix $\mathbf{F}$ is changed by subtracting the information contained in the selected sentence:

$$\mathbf{F}^{(it+1)} = \mathbf{F}^{(it)} - \frac{\mathbf{f}_{best} \cdot \mathbf{f}_{best}^T}{|\mathbf{f}_{best}|^2} \cdot \mathbf{F}^{(it)}$$

The vector lengths of similar sentences are thus decreased, which avoids selecting the same/similar sentences. We call this a *redundancy filter*. After this subtraction, the process continues with the sentence which has the largest co-occurrence sentence score computed on the updated matrix $\mathbf{F}^1$ (the first update of the original matrix $\mathbf{F}^0$). The process is repeated until all the sentences of the document(s) are annotated with their co-occurrence sentence score.

Since it is unsupervised, in our work this method was applied to both the source language texts and the translated texts.

4.2 Profile-based ranking

The supervised profile-based ranking algorithm by (Pouliquen et al., 2003) was proposed for addressing the multi-label categorization problem using the Eurovoc thesaurus[3]. Models for thousands of categories were trained using only positive samples for each category. The training process consisted in identifying a list of representative words and associating to each of them a log-likelihood weight, using the training set as the reference corpus. A new document was represented as a vector of words with their frequency in the document. The most appropriate categories for the new document were found by ranking the category vector representations (the *profiles*) according to their cosine similarity to the vector representation of the new document.

In this paper we are primarily interested in the ranking of sentences, as opposed to the ranking of categories. Since we know beforehand which category (a topic of interest) a document belongs to, a profile vector is created for that category using human labeled data. The cosine similarity for each sentence in the document and the category vector is computed and all the sentences are ranked according their cosine value.

In our work this method was applied to the source language sentences only.

5 Experimental settings

5.1 Corpora

Relevance ranking training The profile-based method (Section 4.2) is trained using $1,000$ French news documents for each of our four topics of interest. These documents were selected using an in-house news categorization system (Steinberger et al., 2009), where category definitions are created by humans. Articles are said to fall into a given category if they satisfy the category definition, which consists of Boolean operators with optional vicinity operators and wild cards. Alternative classifiers can also be trained using the Eurovoc human labeled multi-lingual resource.

Quality estimation training To train the regression algorithm for the quality estimation model we use the French-English corpus created in (Specia, 2011), which is freely available[4]. This corpus contains $2,525$ French news sentences from the WMT *news-test2009* dataset and their translations into English using a statistical machine translation system built from the Moses toolkit[5]. These sentences were scored by a human translator according to the effort necessary to correct them: 1 = requires complete retranslation; 2 = requires some retranslation; 3 = very little

[2]The degree of importance of each 'latent' topic is given by the singular values and the optimal number of latent topics (i.e., dimensions) r can be tuned on some development data.

[3]Eurovoc.europa.eu/

[4]www.dcs.shef.ac.uk/~lucia/resources.html

[5]www.statmt.org/wmt10/

post editing needed; 4 = fit for purpose. An average human score of 2.83 was reported.

Evaluation corpus To evaluate the performance of our approach we use the multilingual summary evaluation dataset created by Turchi et al. (2010)[6]. It contains four sets of documents covering four topics: *Israeli-Palestinian conflict* (**IPC**), *Malaria* (**M**), *Genetics* (**G**) and *Science and Society* (**SS**). Each set contains five documents, here in French. All sentences (amounting to 789) in these documents were annotated by four human annotators with binary labels indicating whether or not it is informative to that topic. Therefore, the final score for each sentence is a discrete number ranging from 0 (uninformative) to 4 (very informative). These French sentences were then translated using the same Moses system as in the training set for quality estimation and annotated for quality using the 1-4 scoring scheme. The average human quality scores are shown in Table 2.

5.2 Evaluation metrics for ranking

Our goal is to find the best possible ranking of translated sentences and documents according to their relevance and translation quality. While the ranked sentences/documents could be used for many applications, including cross-lingual summarization, we are interested in a more general ranking approach, and therefore our evaluation is task-independent. We use the following metrics:

Sentence ranking Sentences in the system output and gold standard documents are first ordered according to their combined score for relevance and translation quality (or relevance score only, for the monolingual ranking evaluation, Table 1). We then compute the Spearman's rank correlation coefficient (ρ) between the two rankings. Additionally, inspired by the vBLEUΔ metric (Soricut and Echihabi, 2010), we compute $Avg\Delta$, a metric that measures the relative gain (or loss) in performance obtained from selecting the top $k\%$ sentences ranked according to the predicted scores, as compared to the performance obtained from randomly selecting $k\%$ sentences:

$$Avg\Delta = (Avg_{sys} - Avg_{gold})$$

where Avg_{gold} is the average *gold-standard* score for *all* sentences in the test set (i.e., the approximate score if sentences are randomly taken) and Avg_{sys} is the average *gold-standard* score for the top $k\%$ sentences from the test set ranked according to the predicted (system) scores.

Intuitively, the smaller the k, the higher the upper bound $Avg\Delta$, but the harder the ranking task becomes. Larger values of k should result in smaller values for $Avg\Delta$. For k = 100, $Avg\Delta$ = 0. In this paper we compute $Avg\Delta$ over different values of k: 10, 25 and 50, and consider the arithmetic mean over these values of k as our final metric, $Avg\Delta_{all}$.

Document ranking Likewise in sentence ranking, both gold-standard and system rankings for the documents are compared. Since there are only five documents within each set of documents, Spearman's rank correlation coefficient would not be reliable. We instead evaluate the pairwise rankings of documents using Cohen's Kappa coefficient (κ) (Cohen, 1960), defined as: $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, where $P(A)$ is the proportion of times the gold-standard and system ranking agree on the ranking of a pair of documents and $P(E)$ is the proportion of times they could agree by chance. This probability is empirically computed by observing the frequency of ties, as in (Callison-Burch et al., 2011).

6 Experiments and results

In what follows we show the results of the quality estimation and relevance ranking methods on their own and then we present the results obtained with the combination of these two methods.

6.1 Quality estimation

The performance of the quality estimation method is shown in Table 2. The average regression error is measured using Root Mean Squared Error, RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}$, where N is the number of test sentences, $\widehat{y}$ is the predicted score and y is the actual score for that test sentence. The performance is generally lower than what has been reported in (Specia, 2011) for French-English and similar settings (RMSE = 0.662). The decrease in performance is most likely due to the difference in the text domain of the training and test

[6] langtech.jrc.it/JRC_Resources.html

	G		IPC		M		SS		Macro Av.	
	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ
InvPos	-0.254	-0.088	-0.08	0.006	-0.22	0.012	0.132	0.015	-0.105	-0.013
Length	0.287	0.328	0.322	0.278	**0.75**	**0.541**	0.156	0.113	0.378	0.315
PB 1000	0.312	0.285	0.358	0.321	0.329	0.286	**0.227**	0.072	0.307	0.242
PB 2000	**0.568**	**0.401**	**0.568**	**0.338**	0.385	0.303	0.154	**0.141**	0.419	0.296
PB 5000	0.478	0.249	0.503	0.31	**0.607**	**0.451**	0.046	0.095	0.409	0.271
Co_R_S 25	**0.293**	**0.364**	0.469	0.301	0.544	0.428	0.203	0.244	0.377	0.335
Co_NR_S 2	0.267	0.269	0.388	0.236	0.28	0.389	**0.607**	**0.367**	0.386	0.316
Co_NR_S 5	0.12	0.224	**0.605**	0.3	0.394	0.389	0.412	0.365	0.382	0.32
Co_R_D 25	0.292	0.295	0.53	**0.362**	**0.589**	**0.461**	0.18	0.208	0.398	0.332
Co_R_D 5	0.271	0.263	0.446	0.335	0.546	0.41	0.183	0.296	0.362	0.326
Oracle	1.559	1	1.623	1	1.453	1	1.5	1		
Lower bound	-0.94	-1	-0.898	-1	-0.726	-1	-0.9	-1		

Table 1: Performance of the sentence ranking methods on monolingual data. PB: profile-based ranker; Co: co-occurrence-based ranker; R/NR: Redundancy reduction enabled/disabled; D/S: ranking based on individual documents or sets of documents on the same topic of interest. The *Oracle* values are obtained using the gold-standard ranking, while the *Lower bound* values consider the inverted gold-standard ranking.

Topic	Avg. human score	RMSE
IPC	3.29	0.696
G	3.00	0.755
M	3.14	0.734
SS	2.89	0.712

Table 2: Average human score and regression error of the quality estimation approach.

datasets. The training dataset covers main news stories from September to October 2008, while the test set covers news commentaries on specific topics from 2005 to 2009.

6.2 Monolingual relevance ranking

The performance of the relevance ranking methods on the original, source-language texts is shown in Table 1. For the unsupervised co-occurrence ranking (Co), we run a number of experiments with different settings. We perform a greedy search on the number of dimensions to be used: 1, and 2%, 5%, 10%, 25% or 40% of the total. We run several experiments enabling (R) and disabling (NR) the sentence redundancy filter and on the full set of documents (S) and on a single document (D). We report here the settings that work the best across different topics. For the profile-based ranking (PB), based on our previous experience with this method, we chose to use the following numbers of words defining the profile vector: $1,000$, $2,000$ and $5,000$.

To define the gold-standard scores for the evaluation at sentence level, we use the number of annotators who selected the sentence as relevant (0-4). The results in Table 1 are the average performance for all documents within a set of documents for each topic. They are compared against baselines proposed in (Kennedy and Szpakowicz, 2011):

- Inverse position (*InvPos*): each sentence is associated with the inverse of its position in the document. The ranking of the sentences thus corresponds to their position in the document and the inverse position is used as their relevance score.
- Sentence length (*Length*): each sentence is associated with the number of words that it contains. Longer sentences are deemed more informative.

The proposed baselines are highly competitive, in particular *Length*. This reflects the fact that longer sentences are naturally better candidates to be more informative, simply because they contain more words. Both methods in all settings outperform the *InvPos* ranker. Except for the **M** topic, most settings of the co-occurrence method and at least one setting of the profile-based method outperform *Length* according to $Avg\Delta_{all}$.

The last column of the Table shows that on average (all topics), the profile-based method seems to be slightly better suited for ranking the top 50% documents, with better $Avg\Delta_{all}$, while the co-occurrence-based method seems to be better for producing a global ranking of all sentences in the dataset, with better ρ coefficient. While the performances of the variations of the co-occurrence-based method seem to be highly dependent on the topic of the documents, it can be observed that on

	G		IPC		M		SS		Macro Av.	
	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ	$Avg\Delta_{all}$	ρ
Length	0.593	0.272	0.886	0.259	2.075	*0.512*	0.365	0.089	0.981	0.283
Length_QE	0.853	0.28	1.02	0.258	2.156	0.518	0.5	0.096	1.132	0.288
Co-Tr_R_S 25	0.374	0.177	1.527	0.31	**1.843**	0.398	0.607	0.197	1.087	0.27
Co-Tr_NR_S 5	0.574	0.276	1.284	0.302	0.832	0.341	**1.196**	**0.344**	0.971	0.315
Co-Tr_NR_S 2	**0.945**	**0.282**	1.518	0.242	1.393	0.377	1.174	0.313	1.257	0.303
Co-Tr_R_D 25	0.834	0.217	1.577	**0.323**	1.668	**0.44**	0.99	0.246	1.267	0.306
Co-Tr_R_D 5	0.752	0.238	**1.598**	0.289	1.536	0.341	1.101	0.274	1.246	0.285
PB 1000	0.853	0.262	1.018	0.304	0.726	0.268	**0.657**	0.06	0.814	0.224
PB 2000	**1.78**	**0.386**	1.375	**0.318**	1.19	0.318	0.642	0.12	1.247	0.286
PB 5000	1.455	0.239	**1.589**	0.279	**1.926**	**0.41**	0.06	**0.062**	1.258	0.248
Co_R_S 25	0.728	**0.327**	1.521	0.299	**1.768**	**0.405**	0.665	0.222	1.171	0.314
Co_NR_S 5	0.443	0.198	1.494	0.275	1.262	0.361	0.947	**0.349**	1.037	0.296
Co_NR_S 2	**0.981**	0.241	1.121	0.23	0.944	0.369	**1.383**	0.34	1.108	0.295
Co_R_D 25	0.729	0.262	**2.163**	**0.341**	1.481	0.402	0.68	0.172	1.264	0.294
Co_R_D 5	0.77	0.21	1.326	0.317	1.344	0.384	0.534	0.23	0.994	0.286
Oracle	5.249	1	4.109	1	3.854	1	3.707	1		
Lower bound	-2.859	-1	-2.335	-1	-1.844	-1	-2.097	-1		

Table 3: Performance of the approaches combining informativeness and quality estimation for sentence ranking. Co-Tr: co-occurrence-based ranker applied directly to translated sentences; PB: profile-based ranker combined with quality estimates, Co: co-occurrence-based ranker applied to source texts and combined with quality estimates. R/NR and D/S as in Table 1.

average across different topics all these variations perform similarly.

We used the same methods - except the *InvPos* baseline, which clearly performs very poorly - and settings to assess the ranking of translated documents.

6.3 Relevance ranking for translated texts

We combine the translation quality and sentence ranking scores for each translated sentence t_i by taking their product:

$$score(t_i) = relevance(s_i) \times quality(t_i)$$

where $relevance(s_i)$ is given by either the co-occurrence (Co) or profile-based (PB) methods applied to the source language sentence s_i, and $quality(t_i)$ is given by the quality estimation method applied to the translation of s_i.

This is done for both the gold-standard annotation and the systems' predictions. The ranges of these two values are different, but this difference is not relevant, since we are only interested in the ranking of the sentences, as opposed to their absolute scores.

Using the product for combining scores is however not ideal: a translation with very low quality but high relevance can receive comparable scores as translations with high quality but low relevance. We have also experimented with using quality estimates as a filter for the relevance rankings. In other words, setting a threshold on the translation quality scores below which a translated sentence is ranked at the bottom of the list even if its corresponding source is highly relevant. This strategy however was strongly affected by the choice of the threshold and resulted in generally poorer performance. Due to space constraints, we only present the results using the product of the two scores.

In the first set of experiments we evaluate the ability of our approach to rank translated sentences within a document. We combine the quality and the relevance scores at sentence level as explained above. As an alternative approach, we apply the unsupervised co-occurrence-based method (Co-Tr) to directly estimate the relevance of the translated text without any quality filtering. In this case, $score(t_i) = relevance(t_i)$. This approach does not explicitly address translation performance. Nevertheless, it can account for some translation problems implicitly, particularly words left untranslated or translated incorrectly. In all cases, the evaluation is performed comparing the system outputs against the combined (product) gold-standard. Results are shown in Table 3. The *Length* baseline is the same as in the monolingual setting and does not include the quality estimation filter. It is also compared against the combined gold-standard.

It is interesting to note that the quality estimation has a positive impact even for the baseline *Length_QE*, confirming that long sentences are often badly translated. The performance of most set-

tings of the co-occurrence and profile-based methods outperform the baselines, except for the **M** topic, as in the monolingual experiments. On average, the co-occurrence method on translated and source data provides better performance than the profile-based method in terms of ρ, while all methods are comparable according to $Avg\Delta_{all}$. This seems to indicate that the profile-based is good at ranking good quality informative sentences, but fails at ranking informative but poorly translated sentences. A possible reason is that it scores each sentence independently from the others and relies on the quality of the training data.

The best settings of the co-occurrence-based method applied to the source language texts outperform the best settings of the same method applied to translated texts. This is more evident in terms of $Avg\Delta$, as opposed to ρ. This seems to indicate that the combination strategy based on the product of the translation quality and relevance scores may not be the most appropriate for fine-grained ranking. Although the monolingual (Table 1) and cross-lingual (Table 3) results are not directly comparable because of their different upper and lower bounds (due to the different gold-standard values in each of these experiments), we can note similar trends with respect to the two ranking methods, Co and PB.

In the second set of experiments we assess the task of ranking documents within a set of documents on the same topic. To produce a unique score for each document, the sentence scores are scaled into $[0, 1]$ and averaged. Documents are then ranked according their average values within their respective groups. The same process is performed using the gold-standard scores and the κ is computed, as shown in Table 4.

The best scores of the proposed approaches vary from moderate to substantial. For the **G**, **IPC** and **M** topics, the best settings of the co-occurrence-based method on the source language outperform the baselines and is superior or equal the other methods. For the **SS** topic, the *Length* baseline is the best method. The co-occurrence method applied directly on the translated sentences is often as good as the two proposed methods that use the source language data. The co-occurrence methods on translated text can in fact be better for heterogeneous sets of documents such as **M**, but in general the usage of source language text can be beneficial.

Overall, the experiments in this paper show significant variations in performance for different methods and settings of the same method over different topics. We believe this is mostly due to the differences in the level of homogeneity of the documents within each topic. Nevertheless, if we consider only the average results over the four topics, we find that most methods/settings perform similarly. This average result however hides significant differences between the methods/settings and opens the way for future research into a better understanding of how to select the best methods and settings for different types of corpora.

	G	IPC	M	SS
Length	0.4	0.4	0.2	0.8
Length_QE	0.4	0.6	0.2	0.6
Co-Tr_R_S 25	0.6	**0.4**	**0.4**	**0.6**
Co-Tr_NR_S 5	0.8	0.0	**0.4**	0.4
Co-Tr_NR_S 2	0.4	0.0	**0.4**	0.2
Co-Tr_R_D 25	0.6	0.0	**0.4**	0.2
Co-Tr_R_D 5	**1.0**	**0.4**	0.0	0.4
PB 1000	**0.8**	0.4	**0.2**	0.0
PB 2000	**0.8**	**0.6**	0.0	0.0
PB 5000	**0.8**	**0.6**	**0.2**	**0.2**
Co_R_S 25	0.8	0.0	0.2	**0.4**
Co_NR_S 5	0.6	**0.4**	0.2	0.0
Co_NR_S 2	0.6	0.2	0.2	**0.4**
Co_R_D 25	0.2	0.0	**0.4**	0.2
Co_R_D 5	**1.0**	0.0	0.0	0.0

Table 4: Kappa coefficient of the various approaches combining informativeness and quality estimation for document ranking.

7 Conclusions and future work

We have proposed combining source relevance information and translation quality estimates to rank translated sentences and documents within groups of texts on the same topic. The approach has shown promising results and it is potentially useful in different scenarios. These include applications where large numbers of documents with redundant information are clustered together according to certain criteria, for example, news on a given topic in media monitoring and news analysis applications, or reviews on a given product/service, and then machine translated to be published in other languages. In this scenario, it would be wise to select for publication only a subset of those documents whose translations are both relevant and of good quality. Additionally, the identification of relevant and high-quality sentences in documents can be used to highlight portions of a document that can be relied upon for gisting purposes, especially in cases where the reader does not have

access to the source document.

In future work, we plan to investigate better ways of combining the translation quality and relevance scores, as well as further investigate the effects of methods and settings on different topics.

References

Bach, N., F. Huang, and Y. Al-Onaizan. 2011. Goodness: A Method for Measuring Machine Translation Confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 211–219, Portland.

Boudin, F., S. Huet, and J.M. Torres-Moreno. 2010. A graph-based approach to cross-language multi-document summarization. *Research journal on Computer science and computer engineering with applications (Polibits)*, 1:21–24.

Callison-Burch, C., P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Carlson, L., J.M. Conroy, D. Marcu, D.P. O'Leary, M.E. Okurowski, A. Taylor, and W. Wong. 2001. An empirical study of the relation between abstracts, extracts, and the discourse structure of texts. In *Proceedings of Document Understanding Conference*.

Chang, C. and C. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27–27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Goldstein, J., M. Kantrowitz, V. Mittal, and J.G. Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. *Computer Science Department*, page 347.

Gong, Y. and L. Xin. 2002. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of ACM SIGIR*, New Orleans, US.

He, Y., Y. Ma, J. van Genabith, and A. Way. 2010. 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.

Kennedy, A. and S. Szpakowicz. 2011. Evaluation of a sentence ranker for text summarization based on rogets thesaurus. In *Text, Speech and Dialogue*, pages 101–108. Springer.

Mihalcea, R. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 20.

Pouliquen, B., R. Steinberger, and C. Ignat. 2003. Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proceedings of the workshop Ontologies and Information Extraction at the EUROLAN'2003, Bucharest, Romania.*

Savoy, J. and L. Dolamic. 2009. How effective is google's translation service in search? *Communications of the ACM*, 52(10):139–143.

Soricut, R. and A. Echihabi. 2010. 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.

Specia, L. and A. Farzindar. 2010. 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*, Denver, Colorado.

Specia, L. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven, Belgium.

Steinberger, J. and K. Ježek. 2004. Text summarization and singular value decomposition. In *Proceedings of the 3rd ADVIS conference*, Izmir, Turkey.

Steinberger, J. and K. Ježek. 2009. Update summarization based on novel topic distribution. In *Proceedings of the 9th ACM Symposium on Document Engineering, Munich, Germany*.

Steinberger, R., B. Pouliquen, and E. Van der Goot. 2009. An introduction to the europe media monitor family of applications. In *Information Access in a Multilingual World-Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR 2009)*, pages 1–8.

Turchi, M., J. Steinberger, M. Kabadjov, and R. Steinberger. 2010. Using parallel corpora for multilingual (multi-document) summarisation evaluation. *Multilingual and Multimodal Information Access Evaluation*, pages 52–63.

Wan, X., H. Li, and J. Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926.

Yeh, J.Y., H.R. Ke, and W.P. Yang. 2008. iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. *Expert Systems with Applications*, 35(3):1451–1462.