ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D1.3
Evaluation and elaboration of metrics

Version No. 1.0
31/12/2011

Document Information

Deliverable number:	D1.3
Deliverable title:	Evaluation and Elaboration of Metrics
Due date of deliverable:	31/12/2011
Actual submission date of deliverable:	30/12/2011
Main Author(s):	Fangzhong Su, Bogdan Babych, Monica Paramita, Rob Gaizauskas
Participants:	CTS, Tilde, USFD, ILSP, FFZG, RACAI
Internal reviewer:	Tilde, DFKI
Workpackage:	WP1
Workpackage title:	Comparability metrics
Workpackage leader:	CTS
Dissemination Level:	**PU**: public
Version:	V1.0
Keywords:	Comparable corpora, machine translation, parallel phrase extraction, comparability metric, lexical mapping

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
V0.1	07/11/2011	draft	CTS	skeleton	Skeleton
V0.2	31/11/2011	draft	CTS	draft	First internal draft
V0.3	02/12/2011	draft	CTS	draft	For internal review at TILDE
V0.4	15/12/2011	draft	CTS	Updated based on TILDE comments	For internal review at DFKI
V0.5	25/12/2011	draft	CTS	Updated based on DFKI comments	Deliverable finished and published
V1.0	31/12/2011	final	Tilde	Final updates	Submitted to PO

EXECUTIVE SUMMARY

This document describes two different comparability metrics to measure the comparability of bilingual texts: a machine translated based metric and a lexical mapping based metric. It also presents experiments to confirm the reliability of the proposed metrics in determining comparability levels.

In order to further investigate the applicability of the metrics in other NLP tasks, the metrics are then combined with the task of parallel phrase extraction from comparable corpora. The experimental results show that both the metrics can help to select more comparable document pairs to improve the performance of parallel phrase extraction.

Table of Contents

Abbreviations

Table 1 Abbreviations and acronyms

Abbreviation	Term/definition
MT	**M**achine **T**ranslation
ACCURAT	**A**nalysis and Evaluation of **C**omparable **C**orpora for **U**nder **R**esourced **A**rea of Machine **T**ranslation
ICC	**I**nitial **C**omparable **C**orpora in ACCURAT
USFD	The comparable corpus collected by University of Sheffield in ACCURAT
SMT	**S**tatistical **M**achine **T**ranslation
WWW	**W**orld **W**ide **W**eb
OOV	**O**ut-**O**f-**V**ocabulary
NLP	**N**atural **L**anguage **P**rocessing

1. Introduction

1.1. Background

Parallel corpora are extensively exploited in different ways in machine translation, as various useful information can be mined from them to improve machine translation performance. However, given the difficulties (e.g., hard criterium of parallelity) in collecting parallel corpora, especially for under-resourced languages and domains, in recent years, the use of cross-lingual comparable corpora has attracted considerable attention in the MT community. In comparison to the collection of parallel corpora, the rich available resources on the World Wide Web (WWW) allow people to relatively easily build comparable corpora from them.

Most of the applications of comparable corpora focus on the detection of translation equivalence from them. For example, comparable corpora have been successfully used for the tasks of bilingual lexicon extraction (Rapp 1995, Rapp 1999, Yu et al. 2010, Prochasson and Fung 2011; Morin et al. 2007, Li and Gaussier, 2010, and Li et al. 2011), parallel phrase extraction (Munteanu and Marcu, 2006), and parallel sentence extraction (Munteanu and Marcu, 2005).

1.2 Purpose of this work

Successful detection of translation equivalents from comparable corpora very much depends on the quality of these corpora, specifically – on the degree of their textual equivalence and successful alignment on various text units. Therefore, the goal of this work is to provide comparability metrics which can reliably identify comparable documents from raw corpora collected by crawling the web, and characterize the degree of their similarity, which enriches comparable corpora with the document alignment information, filters out documents that are not useful and eventually leads to extraction of good-quality translation equivalents from the corpora.

To achieve this goal, we need to define a scale to assess comparability qualitatively, metrics to measure comparability quantitatively, and the sources to get comparable corpora from. The term “comparability”, which is the key concept for this task, applies to the level of corpora, documents and sub-document units. However, so far there is no widely accepted definition of comparability, even though this concept has been frequently used informally, to characterize the overlap in the subject domain or genre of the compared documents. Different definitions of comparability might be given to suit various NLP tasks (see Deliverable 1.1 for an overview about the existing definition of comparability).

Therefore, for the purposes of our study, we can directly characterize comparability by how useful comparable corpora are for the task of detecting translation equivalents in them, and ultimately to machine translation. Comparability measures can be applied on different granularities, such as corpus level, document level and sentence level.

In this work, we focus on document-level comparability, and use three broad categories for qualitative definition of comparability levels:

- Parallel documents are traditional parallel texts that are translations of each other.
- Strongly-comparable documents are independently-written texts in different languages that talk about the same event or subject (e.g., linked articles in Wikipedia about the same topic). These documents can be aligned on the document level on the basis of their origin.

- Weakly-comparable documents are texts in the same narrow domain which describe different events, e.g., customer reviews about hotel and restaurant in London. These texts do not have an independent alignment across languages.

In addition, if pairs of texts are drawn at random from a pair of very large collections of texts (e.g. the web) in the two languages, they are seen as "non-comparable". A more detailed description about the definition of comparability and different comparability levels can be found in Deliverable 1.1 (Babych(a) et al. 2010).

Previously we have proposed a keyword based metric for measuring document level comparability, which is presented in Deliverable 1.2 (Babych(b) et al. 2010). Generally, given a corpus containing both source language documents and target language documents, keyword based metric involves the following steps.

(1) Generate bilingual dictionaries via word alignment process on large-scale parallel corpora.
(2) Compute the absolute frequency and relative frequency of the words in each document.
(3) Generate word frequency list for the whole corpus.
(4) Given the word frequency information, extract keywords from each document by using log-likelihood co-occurrence statistics. A document is then represented by a keyword vector.
(5) Translate keyword vectors in source language into target language by looking up the bilingual dictionaries.
(6) Apply cosine similarity measure to the keyword vectors to compute comparability scores for each document pairs.

More details about the keyword based comparability metric and the evaluation about the metric reliability can be referred in Deliverable 1.2.

In this report, we describe another two different comparability metrics. One is a machine translation based metric, which uses statistical machine translation (SMT) system for document translation and then explores various information. The other one is a lexical overlapping based metric, which uses a bilingual dictionary automatically generated from large-scale parallel corpora for lexical mapping and then compute the comparability strength by cosine similarity measure. We then perform experiments to evaluate the reliability of the proposed metrics, and the experimental results show that the metrics can effectively reflect the comparability levels of document pairs. Furthermore, in order to investigate the usability of the metrics, we also measure their impact on the task of parallel phrase extraction from comparable corpora. It turns out that, higher comparability scores produced from the metrics always lead to more number of parallel phrases extracted from the comparable documents.

Note that both the machine translation based metric and the lexical mapping based metric are unsupervised approaches in essence. In this report, we also further validate the usefulness of features in determining comparability levels by applying 10-fold cross validation (supervised manner) on a manually annotated dataset of Wikipedia documents. In addition, previously in WP3, the project partners have implemented a tool called CNRT (Comparable News

Retrieval Tool) for gathering comparable news texts (see Deliverable 3.4 for details), here we also present the evaluation of this tool on a manually annotated dataset of news texts[1].

1.3 Roadmap

The rest of this report is organized as follows. Section 2 introduces the related work on comparability measure. Section 3 describes the comparability metric based on machine translation and Section 4 introduces the metric using lexical mapping. Experiment and evaluation about the reliability of the proposed metrics are presented in Section 5. In Section 6, we further explore the applicability of the proposed metrics by investigating the impact of comparability metrics to the task of parallel phrase extraction from comparable corpora. In Section 7, we discuss both the advantages and existing problems within the proposed metrics, and point out some possible solutions to help improve the performance of the metrics. In Section 8, we describe the supervised comparability metric for Wikipedia documents, and in Section 9, we present the evaluation of the tool for automatically gathering comparable news text. Finally, in Section 10, we summarize our work in the design of comparability metrics and point out some future work.

1 The evaluation of the tool for gathering comparable news texts has been finished very recently. Given that the information used in this tool is also helpful for comparability metric design, thus we include the evaluation in this report.

2. Related work

Most of the work that uses comparable corpora in NLP applications (e.g., translation equivalence extraction) usually assumes that the corpora they use are reliably comparable. For example, it is common that people crawl data from Wikipedia and see them as comparable corpora. This is because Wikipedia articles have tags linking to articles on the same topics in other languages or tags identifying the domains. By using the tag information, articles that are relevant about the same topic or domain can be collected as comparable corpus. In this case, their focus is on the design of various efficient extraction algorithms but not the comparability measure of corpora. As a result, although data mining in comparable corpora become more and more popular, there is only a few work tackling comparability measure in comparable corpora. A comprehensive introduction about the past research in comparability measure has been presented in Deliverable 1.1. To avoid duplicate description, in this section we will only give a brief review about other work which is not introduced in Deliverable 1.1.

Li and Gaussier (2010, 2011) propose a comparability metric which can be applied at both document level and corpus level and use it as a measure to help select more comparable texts from other external sources into the original corpora. Using the improved comparable corpora, they then perform bilingual lexicon extraction tasks. The idea of their comparability metric is that, given a bilingual dictionary (e.g., EN–FR), it measures the proportion of words in source language corpus translated in the target language corpus by looking up the bilingual dictionary. Both directions (such as translation from EN -> FR and FR -> EN) are measured and then summed up to make the metric symmetrical. They test the proposed approach on English-French comparable corpora, since both English and French are rich-sourced languages and a reliable bilingual dictionary for English and French is available. However, this is not the case for most of the ACCURAT language pairs, as it aims at under-resourced languages and narrow domains and a relatively complete and reliable bilingual dictionary is not available for most of the ACCURAT language pairs.

Munteanu and Marcu (2005, 2006) focus on extracting parallel sub-sentences and sentences from comparable corpora but not on comparability metric design. However, they select more comparable document pairs in an information retrieval based manner by using the publically available toolkit called Lemur (available at http://www.lemurproject.org/). Each document (denoted by D_i) in the source language is translated into the target language and input as a query, it is then compared to all the original documents in target language, and the tool returns the top-n documents to pair with D_i. The automatically generated document pairs are thus comparable and serve as input for the tasks of parallel sentence and sub-sentence extraction.

3. Machine translation based metric

To measure the comparability strength of two documents in different languages (A and B), we need to translate or map lexical items from the text in Language A into Language B, so that we can compare them within the same Language B. Usually this mapping is done by using bilingual dictionaries (Rapp 1999, Li and Gausier, 2010; Prochasson and Fung, 2011) or existing machine translation tools. In this section, we present the comparability metric which uses a machine translation system for text translation and then combines several different types of information in an ensemble manner.

3.1. Statistical MT system for document translation

We use the existing machine translation tools (Google Translator and Microsoft Bing Translator, in the form of open API interface[2]) for document translation. Google translator and Bing translator are state-of-the-art MT systems. In comparison to other MT systems, they train their models based on a much larger-scale data collection and take advantage of their powerful running environment, thus can provide efficient and high-quality translation results. For example, in their MT systems, it is quite rare for them to meet out-of-vocabulary (OOV) problems, which is quite often in a common MT system in which the used training data can not have such broad word coverage.

If the language pair contains a well-resourced and an under-resourced language, (e.g., English-Lithuanian), we usually translate the documents from the under-resourced language into the better-resourced language (English). In case that both languages are under-resourced languages (e.g., EL-RO), their documents are both translated into English. This allows us to apply various available NLP tools (e.g., POS tagging, word stemming and lemmatization, and named entity recognition) on the side of the well-resourced languages and gives additional useful information for comparability metric.

3.2. Mining useful features

For our comparability metric, we extract the following features from each of the compared document pairs.

- Lexical features: Lemmatized bag-of-word representation of each document after stop-word filtering. Obviously, the proportion of overlapped lexical information in two documents is the key factor in measuring their comparability. Higher proportion of lexical overlap indicates that two documents are more comparable. We apply cosine similarity measure to the lexical feature vectors and obtain the lexical similarity score (denoted by W_L) for each compared pair of documents.
- Structure similarity: We approximate it by the number of content words (adjectives, adverbs, nouns, verbs and proper nouns) and the number of sentences in each document, denoted by C_D and S_D respectively. The intuition is that if two

2 Both Google and Microsoft provide free APIs to access their translation service. However, we note that the service of Google Translator API is shut off completely on December 1, 2011. Also, Microsoft has released a new Microsoft Translator API through Windows Azure Marketplace (*www.microsoft.com/Windows*Azure) in September, 2011. But in the previous released APIs, as long as the users send translation requests less than 50 times per minute and the length of each translation request is less than 10000 characters, their APIs can provide free, efficient and high-quality text translation. Since the proposal and experiment of using Google and Bing Translation APIs for document translation were done before the changes in their translation service, here in this report we still present the approach by making use of Bing translation. But in the future, as our project partner DFKI has used large-scale parallel corpora from JRC-Acquis to train baseline MT systems with the Moses SMT toolkit, we will use these baseline MT systems to perform document translation locally.

documents are parallel or strongly-comparable, their number of content words and their document lengths should be similar. For example, two articles might be talking about the same subject, such as "Manchester United", but the longer one is more likely to cover more information which is not included in the shorter one. Thus, the structure similarity (denoted by W_S) of two documents D1 and D2 is defined as bellow.

$$W_S=0.5*(C_{D1}/C_{D2})+ 0.5*(S_{D1}/S_{D2}),$$

suppose that $C_{D1}<=C_{D2}$, and $S_{D1}<=S_{D2}$ (switch C_{D1} and C_{D2} if $C_{D1}>C_{D2}$, and S_{D1} and S_{D2}, if $S_{D1}>S_{D2}$). Note that for structure similarity, we assign equal weight (0.5) to both content word numbers and sentence numbers.

- Keyword features: Top-20 words (ranked by TFIDF weight) of each document. The idea is that TFIDF measures the weight of terms in the documents, thus it can help select more informative words (keywords) from document. If any two documents share more keywords, they should be more comparable (actually this is also the main idea of keyword based metric). Cosine similarity measure is applied to capture the keyword similarity (denoted by W_K) of each document pair.
- Named Entity features: Named entities identified in each document. If more named entities are the same in two documents, these documents are very likely to talk about the same event or subject and thus should be more comparable. Again, cosine similarity is applied to measure the closeness between named entity vectors (denoted by W_N) in each compared document pair.

3.3. Formulation of the metric

After obtaining the four individual comparability scores (W_L, W_S, W_K, and W_N) for lexical feature, structure feature, keywords and named entities, we apply a weighted average strategy to combine these different types of scores in the comparability metric. Specifically, in the metric, each individual score is associated with a constant weight, indicating the relative confidence (or importance) of the corresponding type of score. Thus, the overall comparability score (denoted by *SC*) of a document pair is computed as below:

$$SC= \alpha *W_L+ \beta *W_S+ \gamma *W_K+ \delta * W_N$$

where α, β, γ, and δ ∈[0, 1], and α+β+γ+δ=1. SC should be a value between 0 and 1, and larger SC value indicates higher comparability level. Overall, in the experiment, for the weight of each type of comparability scores, we assign 0.5 for lexical features, 0.2 to keyword feature and named entity feature, and 0.1 to structure features. The assignment of feature weight is based on an assumption that, lexical feature can best characterize the comparability level of document pairs, while keyword and named entity features are also better indicators of comparability than the simple document length information.

4. Lexical mapping based metric

Using machine translation systems (especially the state-of-the-art MT systems) for text translation can provide good-quality translation results for metric design. However, this process might be time-consuming (the translation models need to explore a large search space, compute and compare various possible translation candidates to find the best translation), especially in the scenario that the comparability metric is used to select more comparable corpora from large-scale web crawled raw data. Thus, in order to speed up the text translation process even at the cost of producing worse translation quality than that of machine translation, we also design another metric which uses lexical mapping for text translation and then measures the proportion of overlapping lexical information between a document pair.

4.1. Automatic generation of bilingual dictionary

The goal of ACCURAT project is to investigate how comparable corpora can compensate for the lack of sufficient linguistic resources to improve MT quality for under-resourced languages and narrow-domains, hence there are very limited resources available for ACCURAT languages. For the purpose of measuring lexical overlapping, it is straightforward that we expect that a bilingual dictionary with good word coverage can be used to check the mapping. However, unlike the language pairs in which both languages are rich-resourced (e.g., English and French) and machine-readable bilingual dictionaries are easy to obtain, for most of the ACCURAT language pairs bilingual dictionaries are small and in many cases not publically available. Fortunately, we can use word alignments to construct a bilingual dictionary. Word alignment, e.g., using GIZA++ (Och and Ney 2000, Och and Ney 2003), are applied to parallel corpora in SMT to create translation models and include lexical information about the source and target texts. Inspired by this, in this work, we construct the bilingual dictionaries by using the word alignment result from GIZA++.

Specifically, a bilingual dictionary is generated as below. We extract parallel corpora for ACCURAT language pairs from JRC-Acquis corpora[3]. Then GIZA++ toolkit is used for word alignment process on the extracted parallel corpora. The word alignment results together with the alignment probabilities are then converted into dictionary entries. For example, in the Romanian-English language pair, the alignment result “companie company 0.625” means that the Romanian word “companie” can be translated into (or aligned with) “company” in English with a probability of 0.625. So in the dictionary of Romanian-English (translate Romanian word into English), “company” will be recorded as a translation candidate together with translation probability for the Romanian word “companie”. The translation candidates are ranked by translation probability in descending order. Note that the dictionary collects inflectional form of words, but not base form of words, as it is likely that a reliable word stemming or lemmatization toolkit is not publically available for the under-resourced languages.

Overall, the information about the automatically generated dictionary for each language pair in ACCURAT, including size of parallel corpora, number of sentences, and size of dictionary, is listed in Table 2.

3 JRC-Acquis provides large-scale parallel corpora for 22 languages, which covers most of the ACCURAT languages except Croatian. More details about JRC-Acquis project can be referred to the project website at http://langtech.jrc.it/JRC-Acquis.html.

Table 2 Bilingual dictionaries generated from JRC-Acquis corpora

	#parallel sentences	Size of parallel corpus (Megabyte)	#entries of dictionary
DE-EN	119025	395	388072
EL-EN	590453	151	221403
EL-RO	306222	174	91658
ET-EN	1088724	336	547102
LT-EN	1155324	363	372430
LV-EN	1088424	349	350519
RO-EN	337406	110	80551
RO-DE	335345	108	79471
SL-EN	1114458	343	314267
LV-LT	1181296	366	320764
RO-LT	361433	113	71667

4.2. Lexical mapping strategy

The lexical mapping process is based on a word-for-word mapping strategy. Given a document in source language, we scan each word in it to check if the word occurs in the dictionary. If so, we select their translation candidates from the dictionary. The translation candidate selection is based on the translation probabilities (alignment probability from GIZA++). If there is only one translation candidate for the source language word *W* in the dictionary, it is returned as the mapping result of W in target language. Suppose that there are more than one translation candidate for the source language word *W*, if the translation probability is higher than 0.3 for the first candidate and lower than 0.1 for the second candidate, only the first candidate is kept as the corresponding mapping of *W*, otherwise the top two candidates are retained. The reason for setting different scales in determining the number of retained translation candidates is that, from the manual inspection on the word alignment results from GIZA++, we find that if the alignment probability is higher than 0.3 it is more reliable; and if it is lower than 0.1, the alignment result is less accurate.

If the source language words do not occur in the dictionary, they will be omitted from the translation process. Thus, by doing this word-for-word mapping, the documents in source language are translated into target language quickly, even at the cost that much important information such as word order, grammar, syntactic information, named entities (as many named entities are not collected in the dictionary) and out-of-vocabulary words is lost in the translated text.

4.3. Formulation of the metric

For non-English and English language pair, after mapping the non-English documents into English, we apply stop-word filtering and word lemmatization process and convert texts into

feature vectors (bag-of-word representation). If both sides of the language pair are not English, after mapping the source language document into target language, we only apply stop-word filtering (the stop-words lists for ACCURAT languages are provided by project partners)[4] on target language documents. Again, cosine similarity measure is applied to determine the comparability strength at the document level.

4 In the future work we will incorporate word lemmatization for non-English languages.

5. Experiments and Evaluation

To investigate the reliability of the proposed comparability metrics, we use the initial comparable corpora (ICC) collected in ACCURAT project for experiments. A detailed description about ICC corpora is given in Deliverable 3.1 (Giouli el al. 2010). ICC contains cross-lingual comparable corpora for under-resourced languages in international news, sports, administration, travel, software, and medicine domains, and comparable corpora in narrow subject domains (e.g., automotive, software, and medicine) for the English-German language pair. A subset of ICC has been annotated at the document level (document pairs) for comparability levels defined in Section 1.2 (parallel, strongly-comparable, weakly-comparable). The annotation was done manually for all language pairs, except English and German. Hence, we use this subset as gold standard, and perform the experiments on 9 language pairs[5]: German-English (DE-EN), Greek-English (EL-EN), Estonian-English (ET-EN), Lithuanian-English (LT-EN), Latvian-English (LV-EN), Romanian-English (RO-EN), Slovenian-English (SL-EN), Greek-Romanian (EL-RO) and Romanian-German (RO-EN).

We adopt a simple method for evaluation. For each language pair, we compute the average scores for all the document pairs in the same comparability level, and compare them to the corresponding comparability levels. In addition, in order to better reveal the relation between the scores obtained from the proposed metrics and comparability levels, we also measure the correlation between the comparability scores and comparability levels to validate if the comparability scores automatically obtained from the metrics are in line with the gold standard labels of comparability levels.

5.1. Evaluation on MT based metric

The experimental results (number of document pairs in one comparability level and the average comparability score of these document pairs) of machine translation based metric on ICC corpora are presented in Figure 1 and in Table 3.

5 The experiment on English-Croatian language is still in progress as we need to find a solution for the lack of large scale parallel corpora.

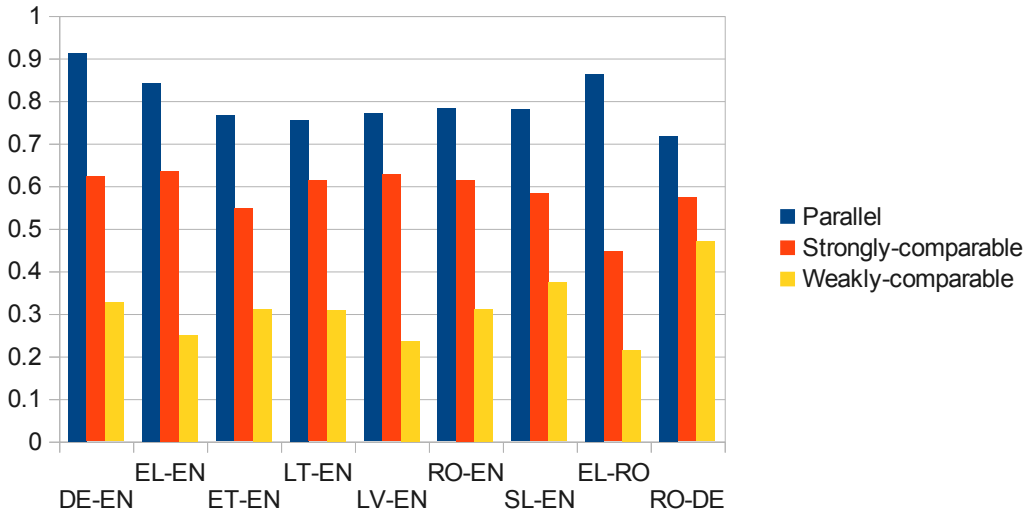

Figure 1 Average comparability scores for each of the comparability levels in ICC (MT based metric)

Table 3 Number of document pairs (top) and average comparability scores (bottom, bold) for different comparability levels in ICC (MT based metric)

Language pair	Overall number of document pairs	Parallel	Strongly-comparable	Weakly-comparable	Correlation
DE-EN	1286	531 **0.912**	715 **0.622**	40 **0.326**	0.999
EL-EN	834	85 **0.841**	400 **0.635**	349 **0.250**	0.985
ET-EN	1648	182 **0.765**	987 **0.547**	479 **0.310**	0.999
LT-EN	1177	347 **0.755**	509 **0.613**	321 **0.308**	0.984
LV-EN	1252	184 **0.770**	558 **0.627**	510 **0.236**	0.966
RO-EN	130	20 **0.782**	42 **0.614**	68 **0.311**	0.987
SL-EN	1795	532 **0.779**	302 **0.582**	961 **0.373**	0.999
EL-RO	485	38 **0.863**	365 **0.446**	82 **0.214**	0.988
RO-DE	167	16	84	67	0.996

Language pair	Overall number of document pairs	Parallel	Strongly-comparable	Weakly-comparable	Correlation
		0.717	**0.573**	**0.469**	

From the average cosine scores for each comparability level presented in Figure 1 and Table 3 we can see that, the scores obtained from the comparability metric can reliably reflect the comparability levels across different languages, as the average scores for higher comparable levels are always significantly larger than that of lower comparable levels, namely SC(parallel)>SC(strongly-comparable)>SC(weakly-comparable). Moreover, the correlation scores also indicate that there is a strong correlation between the comparability scores obtained from the proposed metric and the corresponding comparability level. These results thus confirm that (on the level of average scores for the document collection) the comparability level predicted by our metric corresponds to the independently defined levels of comparability.

5.2. Evaluation on lexical mapping based metric

Still using ICC[6] as gold standard, the experimental results of lexical mapping based metric are shown in Table 4 and Figure 2.

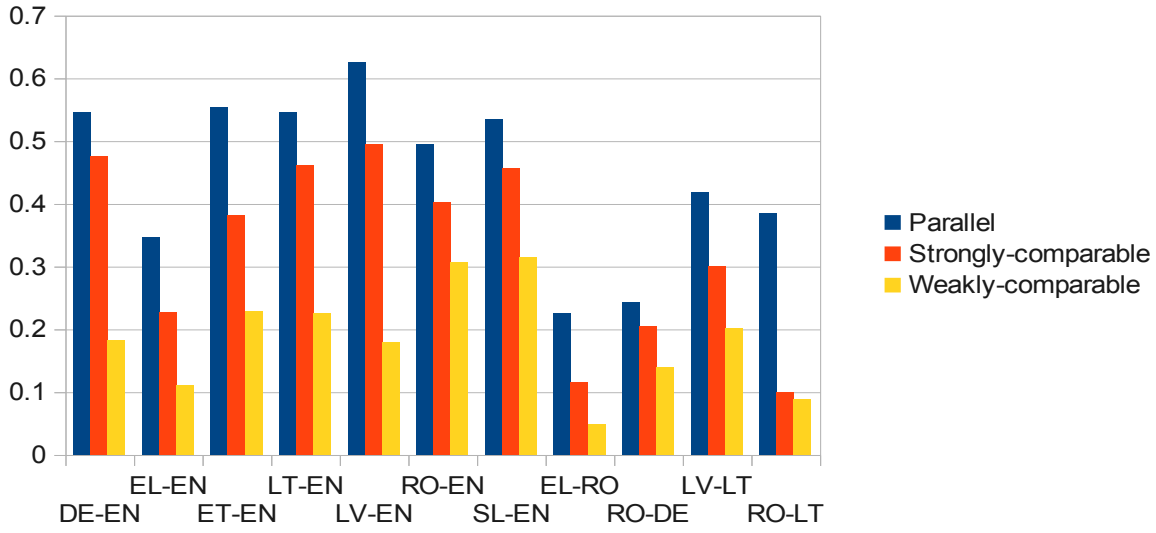

Figure 2 Average comparability scores for each of the comparability levels in ICC (lexical mapping based metric)

6 In the first release of ICC corpora, Latvian-Lithuanian, Romanian-Lithuanian are not included. The MT based metric was applied on this version of ICC, thus we do not evaluate on these two language pairs due to the recent changes in Google and Bing translation services. However, in the lexical mapping based metrics, their evaluation is also included in this report.

Table 4 Number of document pairs (top) and average comparability scores (bottom, bold) for different comparability levels in ICC (lexical mapping based metric)

Language pair	Overall number of document pairs	Parallel	Strongly-comparable	Weakly-comparable	Correlation
DE-EN	1286	531 **0.545**	715 **0.476**	40 **0.182**	0.942
EL-EN	834	85 **0.346**	400 **0.227**	349 **0.111**	0.999
ET-EN	1648	182 **0.553**	987 **0.381**	479 **0.228**	0.999
LT-EN	1177	347 **0.545**	509 **0.461**	321 **0.225**	0.964
LV-EN	1252	184 **0.625**	558 **0.494**	510 **0.179**	0.973
RO-EN	130	20 **0.494**	42 **0.403**	68 **0.307**	0.999
SL-EN	1795	532 **0.535**	302 **0.456**	961 **0.314**	0.987
EL-RO	485	38 **0.225**	365 **0.115**	82 **0.048**	0.990
RO-DE	167	16 **0.243**	84 **0.205**	67 **0.140**	0.989
LV-LT	232	39 **0.418**	114 **0.300**	79 **0.201**	0.999
RO-LT	15812	9123 **0.385**	6688 **0.100**	1 **0.088**	0.883

From the results listed in Table 4 and Figure 2, first we can see that, similar to that in MT-based metric, higher comparability levels also have significantly higher comparability scores generated from the metric. Strong correlation between comparability scores and comparability levels also holds[7] in the simple lexical mapping based metric.

However, we also see that, in each language pair, the average score for each comparability level drops sharply in comparison to that in MT based metric. Particularly even within the same metric, in EL-RO, and RO-DE, we see that the scores obtained from the metric are

7 Correlation score for RO-LT is lower than 0.9, this is because there is only one weakly-comparable document pair in ICC. The average score is computed from only one sample, thus this score can not represent the average comparability score of weakly-comparable document pairs.

rather low, even for parallel document pairs. For example, the scores are 0.225 and 0.243 respectively. The reason is very likely due to the quality of dictionaries. As the size of parallel corpora for these two non-English language pairs is smaller than most of the language pairs in ACCURAT, their word alignment results are thus not as good as the others and the number of the resulting dictionary entries is also smaller. Thus, the low scores for non-English language pairs also indicate that using English as pivot language (e.g., translating both the source and target language texts into English) should be worth a try.

Moreover, we also notice that the average gap between different comparability levels in lexical mapping based metric is also significantly smaller than that of machine translation based metric. For example, 0.099 ***vs.*** 0.214[8] between "parallel" and "strongly-comparable", and 0.165 ***vs.*** 0.274 between "strongly-comparable" and "weakly-comparable". The reason for the decreased scores is that, the translation quality of dictionary-based lexical mapping is worse than MT based approach. Moreover, better translation performance from MT systems also allows detecting more proportion of lexical overlapping and mining more useful information in the translated text for metric design, which in turn can better catch the distinction in different comparability levels. However, even though the comparability scores of dictionary-based metric drop, they still effectively reflect the trend that document pairs with higher comparability level are likely to obtain higher scores. Thus, the lexical mapping based metric is also reliable for predicting comparability levels of document pairs.

8 The average gap between two different comparability levels is calculated as below. We first compute the average score for each comparability level among the 9 language pairs, which are 0.798 (vs. 0.457) for "parallel" level, 0.584 (vs. 0.358) for "strongly-comparable" level, and 0.310 (vs. 0.193) for "weakly-comparable" level in the machine translation based metric (lexical mapping based metric). So in MT based metric, the average gap is 0.798-0.584=0.214 between "parallel" and "strongly-comparable" level, and 0.584-0.310=0.274 between "strongly-comparable" and "weakly-comparable". While in lexical mapping based metric, the corresponding gap is 0.099 and 0.165 respectively.

6. Application of the metrics

The experiments and evaluation in Section 5 confirm the reliability of the proposed metrics. The comparability metrics are thus useful for collecting high-quality comparable corpora, as they can help filter out weakly comparable or non-comparable document pairs from the raw crawled corpora. But are they also useful for other NLP tasks, such as translation equivalence detection from comparable corpora? In this section, we measure the impact of the metrics on parallel phrase extraction from comparable corpora. Our intuition is that, if a document pair is assigned a higher comparability score, it should be more comparable and thus more parallel phrases can be extracted from it.

The algorithm for parallel phrase extraction, which develops the ideas of the algorithm presented in (Munteanu and Marcu, 2006), uses the lexical overlap and the structural matching measures. Overall, taking a list of document pairs in which each pair consists of a source language document and a target language document as input, the extraction algorithm involves the following steps.

- It splits the source and target documents into phrases.
- It then computes the degree of parallelism for each possible pair of phrases by using the bilingual dictionary generated from GIZA++ (base dictionary), and retains all the phrase pairs with a score larger than a predefined threshold.
- GIZA++ is applied to the retained phrase pairs to detect new dictionaries entries, which are then added to the base dictionary.
- Using the augmented dictionary, the algorithm iteratively executes Step 2 and Step 3 for several times (empirically set at 5) and outputs the detected phrase pairs.

A detailed description about the algorithm of parallel phrase extraction is presented in Deliverable 2.6 (Ion et al. 2011).

For the experiment of parallel phrase extraction, we use another dataset (called USFD) collected by our ACCURAT partner at the University of Sheffield (USFD). USFD is a raw comparable corpus crawled from the Web, and much larger than ICC. A description about the way they collect the comparable corpora and the statistics about USFD can be referred to Deliverable 3.4 (Paramita et al. 2011) and Deliverable 3.5 (Aker et al. 2011).

For evaluation, we measure how the metrics affect the performance of the parallel phrase extraction algorithm on 7 language pairs (DE-EN, EL-EN, ET-EN, LT-EN, LV-EN, RO-EN and SL-EN). We first apply our comparability metrics to USFD to assign comparability scores for all the document pairs in USFD. For each language pair, we set three different intervals based on the comparability score (SC). For the MT based metric, the three intervals are (1) 0.1<=SC<0.3, (2) 0.3<=SC<0.5, and (3) SC>=0.5[9]. For the lexical mapping based metric, since its scores are lower than those of the MT based metric for each comparability level, we set the three lower intervals at (1) 0.1<=SC<0.2, (2) 0.2<=SC<0.4, and (3) SC>=0.4. The experiment focuses on counting the number of extracted parallel phrases with parallelism score>=0.4, and computes the average number of extracted phrases per 100000 words (the sum of words in both source language and target language documents) for each interval. The reason that we only take parallel phrase with parallelism score larger or equal to 0.4 is that, from the manual evaluation of the extraction performance carried out by our Romanian partner RACAI (the developer of parallel phrase extraction algorithm), it was

9 We can set other intervals for experiment as well, such as SC>=0.6.

shown that automatically extracted parallel phrase pairs with parallelism score>=0.4 are more reliable.

6.1. Impact of MT based metric

Based on the evaluation setting above, the results which demonstrate the impact of MT based metric in the performance of parallel phrase extraction from comparable documents is presented in Table 5 and Figure 3. For each interval, we list the total number of words of the tested data, the number of extracted parallel phrases with parallelism score>=0.4, and the average number of extracted phrases (in bold) per 100000 words. In addition, Pearson's correlation measure is also applied to measure the correlation between the comparability scores and the number of extracted parallel phrases.

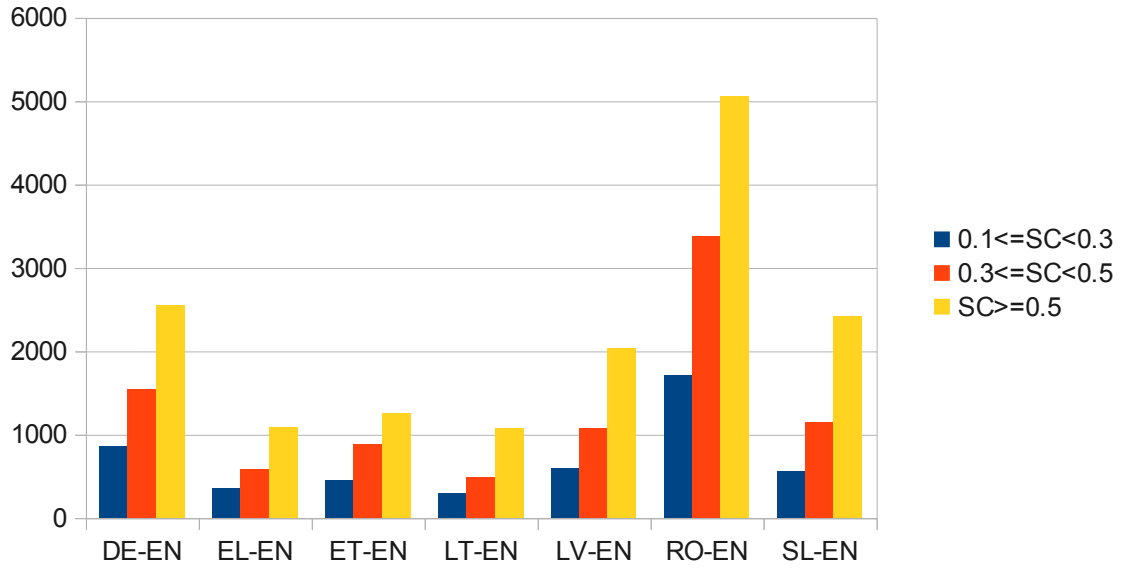

Figure 3 Number of extracted parallel phrases for different intervals for different comparability scores in USFD corpus (MT based metric)

Table 5 Number of extracted parallel phrases for different intervals on USFD (MT based metric)

	0.1<=SC<0.3	0.3<=SC<0.5	SC>=0.5	Pearson's R correlation: average score vs. number of extracted equivalents
DE-EN	5943[10] 687859[11] **861**[12]	674674 10364 **1547**	803044 20413 **2552**	**0.996**

10 The first line in each cell indicates the total number of extracted parallel phrases with parallelism score>=threshold (0.4) in the dataset contains comparable document pairs in the specified interval (e.g., 0.1<=SC<0.3).

11 The second line in each cell indicates the total number of words in the dataset contains comparable document pairs in the specified interval.

12 The third line in bold indicate the average number of extracted phrases per 100000 words. For example, 5943/6.9=861.

	0.1<=SC<0.3	0.3<=SC<0.5	SC>=0.5	Pearson's R correlation: average score vs. number of extracted equivalents
EL-EN	852609 3051 **359**	823545 4739 **578**	845194 9145 **1082**	**0.975**
ET-EN	665142 3002 **448**	633341 5568 **883**	625179 7821 **1251**	**0.999**
LT-EN	691646 2028 **293**	681267 3292 **483**	701401 7505 **1070**	**0.959**
LV-EN	720405 4242 **589**	663713 7075 **1072**	679970 13851 **2037**	**0.982**
RO-EN	748173 12790 **1705**	656046 22298 **3378**	689984 34858 **5052**	**0.999**
SL-EN	624399 3470 **560**	563604 6448 **1151**	580008 14042 **2421**	**0.979**

From Figure 3 and Table 5, we can see that for all the 7 language pairs, based on the average number of extracted aligned phrases, clearly we have interval (3)>(2)>(1). In other words, a higher comparability level always leads to significantly more number of aligned phrases extracted from the comparable documents. In addition, in the same interval, the number of extracted phrases is different across different language pair. For example, in general, many more phrases are extracted in DE-EN and RO-EN, while a smaller number of parallel phrases extracted in EL-EN and LT-EN. The reason might be that, the dictionary and the dataset for test are different for different language pairs.

Pearson's R correlation between the average numeric value of the comparability score and the number of extracted equivalents is very close to 1 for all language pairs, which indicates that the metric results are in line with the performance of equivalent extraction algorithm. So in order to extract more parallel phrases from comparable documents, the comparability metric can be applied beforehand to select more comparable documents, where it is possibly to successfully extract a greater number of translation equivalents.

6.2. Impact of lexical mapping based metric

The results which show the impact of lexical mapping based metric to parallel phrase extraction are presented in Figure 4 and Table 6.

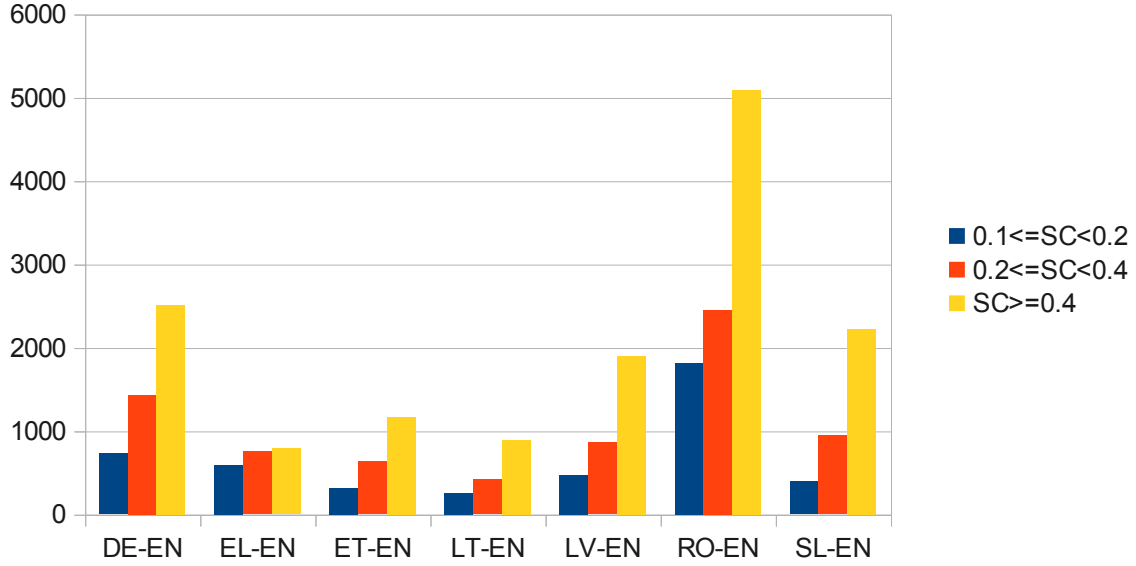

Figure 4 Number of extracted parallel phrases for different intervals for different comparability scores in USFD corpus (Lexical mapping based metric)

Table 6 Number of extracted parallel phrases for different intervals on USFD (Lexical mapping based metric)

	0.1<=SC<0.2	**0.3<=SC<0.4**	**SC>=0.5**	**Pearson's R correlation: average score vs. number of extracted equivalents**
DE-EN	697447 5072 **728**	686804 9855 **1434**	684044 1716 **2510**	**0.993**
EL-EN	862707 5121 **593**	842212 6325 **751**	770292 6093 **791**	**0.946**
ET-EN	712174 2226 **313**	707858 4470 **631**	679933 7928 **1166**	**0.989**
LT-EN	714784 1847 **258**	687213 2878 **419**	823600 7353 **894**	**0.962**
LV-EN	732412 3470 **470**	713394 6123 **859**	720315 13686 **1900**	**0.967**
RO-EN	815186 13276	688033 16855	866632 44095	**0.943**

	0.1<=SC<0.2	0.3<=SC<0.4	SC>=0.5	Pearson's R correlation: average score vs. number of extracted equivalents
	1814	**2450**	**5086**	
SL-EN	669700 2635 **393**	582410 5485 **946**	614089 13630 **2220**	**0.975**

From Table 6 and Figure 4, again we can see that higher comparability scores lead to more parallel phrases extracted from comparable documents. Also, strong correlation between comparability scores obtained from lexical mapping based metric and number of extracted parallel phrases holds. Moreover, although dictionary based metric produces lower comparability scores than MT based metric, they have very similar impact in the task of parallel phrase extraction from comparable documents.

7. Discussion of the metrics

We have proposed three different metrics for measuring comparability at the document level:

- The keyword based metric (presented in detail in Deliverable 1.2)
- The machine translation based metric
- The lexical mapping based metric.

In this section, we will discuss both the advantages and limitations of the proposed metrics, and point out several possible ways to further improve the current metrics.

7.1. Status of the comparability metrics

The keyword based metric allows us to select more informative words from the documents. Therefore, it can reduce the effect of less informative words, which take up a higher proportion than the informative words in a document. Theoretically, if the extracted keywords can well represent the documents, it is promising to measure the comparability by focusing only on the keywords and pruning the large amount of less informative words in the documents.

However, the performance of this metric highly depends on several factors, which are listed as below.

- The keyword extraction algorithm should be reliable. For example, most of the key words that indicate the content or domain of the document should be extracted by the algorithm.
- The metric relies on a bilingual dictionary for keyword translation, whether the dictionary has broad word coverage or not strongly affects the keyword translation quality. However, since the dictionaries for under-resourced language pairs are automatically generated from the available parallel corpora, their quality is not as good as other available machine-readable dictionaries for rich-resourced language pairs. This is because the publically available parallel corpora are either too small or domain specific (for example, Europarl contains only European parliament proceedings, and JRC-acquis focuses on legal documents) and some potential errors occur in the word alignment by using GIZA++, making it hard to generate good and accurate dictionaries with broad word coverage. Therefore, it is possible that translations for a given keyword cannot be found in the dictionary.
- Even when the dictionary provides translation candidates for the keywords in the source language, it is possible that the corresponding translation candidates do not match any keywords in the target language. If two documents are highly comparable, intuitively the translation candidates of keyword in the source language document should at least be similar or relevant to some keywords in the target language document (such as synonyms) even that they are not the same word form. However, in the current metric, we do not further explore the relevance between keywords but only measure the proportion of lexical overlap.

Due to the effect of the above factors, if the metric only focuses on a small number of keywords and only a few overlapped keywords are found, the comparability scores generated from the metric will be very low even that the corresponding document pairs might be highly comparable.

The machine translation based metric can provide good-quality translation results (e.g., by using Google or Bing translator), and various useful information such as named entities is well preserved in the translated text. Also, better translation results allow us to explore more useful features for metric design. The experimental results of the machine translation based metric confirm the effectiveness and reliability in measuring comparability strength at the document level. However, the translation process depends on the availability of powerful MT systems, and it is time-consuming to translate large-scale raw document collections.

The lexical mapping based metric is proposed to address some disadvantages of the keyword based metric although it also encounters some of the problems that exist in keyword based metric. Instead of focusing on keywords only, it performs word-for-word mapping on all the words in the source language document. This reduces the risk of low performance in keyword extraction and low mapping in the extracted keywords. In comparison to machine translation based metric, it is much faster in the document translation process since it adopts a simple word-for-word mapping strategy.

However, the lexical mapping is done via looking up dictionary, thus it also suffers from the word coverage problem in the automatically generated dictionary. In addition, there the translation is done word for word, and words which do not occur in the dictionary are omitted. The word order in the translated texts directly mirrors the structure of the source language, so important information about grammar, morphology, syntactic structure and named entities is lost in the lexical mapping based metric. Thus, apart from the lexical features, it is difficult to mine other useful features due to the relatively low quality of translated results.

7.2. Possible ways for further improvement

Despite the success of the proposed metrics in measuring comparability of comparable document pairs, we also analyse the existing problems in the metrics. Thus, in order to further improve the performance of the current metrics, there are several directions which are worth further exploration.

- Constructing better bilingual dictionaries: This could be achieved by seeking larger-scale parallel corpora which are across more domains, as theoretically more data should help improve the word alignment accuracy and provide more dictionary entries. Also, corpora covering more domains will increase the word coverage, which will address the current problem that the dictionary might contain rich vocabulary in some certain domains only (e.g., legal documents), but very few entries in other domains (e.g., renewable energy). Furthermore, apart from the use of dictionary in comparability metric design, in the overall work flow of ACCURAT project, the bilingual dictionary is also applied in several other tasks in ACCURAT, such as bilingual lexicon extraction and parallel phrase extraction. Thus, various tasks can benefit from the quality enhanced bilingual dictionary.
- Mining word relatedness by exploring distributional similarity of words from large corpora, such as BNC (Lin 1998). The idea is inspired by the fact that the translation results usually contain words which is similar or relevant to (but not exactly the same) the words in the target language documents. For example, obviously “lecture” and “class”, “tariff” and “tax” are different words, but they have high co-occurrence frequency from in BNC . Thus, if the distributional relevance (or closeness) of words (especially keywords) in two document can be better investigated and incorporated in

the design of comparability metric, it should help improve the performance of the metrics.

- Mining word relatedness by using WordNet (Fellbaum 1998). In WordNet, similar words that share the same sense are grouped together by a synset (synonym set). For example, {happy, glad} – eagerly disposed to act or to be of service; "glad to help" , and {phone, headphone, earpiece, earphone} – electro-acoustic transducer for converting electric signals into sounds; it is held over or inserted into the ear; "it was not the typing but the earphones that she disliked". In the specified sense, "happy" and "glad" are synonyms, and "phone", "headphone", "earpiece" and "earphone" are also synonyms. Hence, applying synonym expansion in the metric should help discover more lexical overlap. In addition, apart from synonym expansion, we can also explore other WordNet relations for lexical expansion such as "similar-to" "direct-hypernym" and "direct-hyponym" to detect more words that are strongly relevant with each other, even that they are not synonyms. Also, there are several WordNet similarity packages available on line (for example, WordNet::Similarity is available at http://wn-similarity.sourceforge.net/), which measure the relatedness of Words by using WordNet information. It might be interesting to incorporate these word relatedness information in the comparability metrics to see if it enhances the performance of the metrics.
- ACCURAT project also involves named entity recognition and terminology extraction tasks in under-resourced languages. The metric design might benefit from these tasks. For example, if the performance of these tasks are relatively reliable, their outputs can be incorporated into comparability measure.
- In the current machine translation based metric, it uses Google or Bing translation APIs for document translation which is time-consuming in the communication with the remote MT server (e.g., send translation request to MT server and receive translated results), especially for large amount of raw document translation. Thus, in order to speed up the translation process, we can use translation models provided by our project partner at DFKI, which are trained using Moses SMT toolkit[13] on large-scale parallel corpora (e.g., JRC-Acquis). This will allow us to run the text translation locally to avoid remote communication.

13 Available at http://www.statmt.org/moses/.

8. Comparability Metrics for Wikipedia

Previous sections have explored the use of MT and dictionaries for the development of comparability metrics. In this section we focus on identifying measures of comparability for pairs of Wikipedia articles on the same topic. We begin in Section 8.1 by describing a selection of features which we extract from pairs of Wikipedia documents written in different languages. We combine some of these features into measures of similarity. To analyse the performance of the features and similarity measures we use the evaluation corpus described in Section 8.2. The features and measures are evaluated as a supervised document classification task based on 10-fold cross-validation. The performance of these metrics is discussed in Section 8.3, and finally, conclusion is discussed in Section 8.4.

8.1. Features

In this section, we describe a number of language-independent features that can be extracted from pairs of Wikipedia articles in different languages but on the same topic. We selected features which cover a variety of data types and extraction levels (shown in Table 7). These features can be extracted from Wikipedia without using any linguistic resources, which provides an advantage for under-resourced languages. Four data types were used in the metrics: anchor (links), character n-gram (CNG), word (cognate) overlap and word length (document size). In addition, some of the features are extracted at two different levels:

(1) Document level: features are computed over all text in the Wikipedia article[14].

(2) Sentence level: features are computed on each possible sentence combination between the two paired Wikipedia articles, thus enabling sentences with the highest score to be paired. The final score for a document pair is the average score of all the paired sentences.

Table 7 List of Features and Extraction Level

Features Name	Data Types				Extraction Level		Notes
	Anchor	Char-N-Gram	Word	Word Length	Document	Sentence	
Anchor-Jaccard-Document	√				√		Jaccard similarity
Anchor-Jaccard-Sentence	√					√	Jaccard similarity
CNG-Jaccard-Sentence		√				√	Jaccard similarity
CNG-TFIDF-Document		√			√		Cosine similarity of TF-IDF
CNG-TFIDF-Sentence		√				√	Cosine similarity of TF-IDF
WordOverlap-TFIDF-Document			√		√		Cosine similarity of TF-IDF
WordLength				√	√		Size ratio toward the larger doc

The features for each data types are described in more details below.

[14] Main text in this case represents the full text describing a particular topic. This does not include any tables, images or inter-language links.

8.1.1. Anchor Overlap

Wikipedia documents are enriched with large numbers of anchors: inline links to other Wikipedia articles, which enable readers to refer to other pages in Wikipedia to discover more information. Moreover, Wikipedia also contains inter-language links: links between documents from different languages describing the same topic. By extracting all document titles which are connected using inter-language links, we are able to build a bilingual dictionary. Therefore, given a pair of bilingual documents, we can then translate all anchors from the source language into the target languages, further enabling anchor overlap to be computed using the Jaccard score. This computation is performed on the document level (F1 - **Anchor-Jaccard-Document**) and sentence level (F2 – **Anchor-Jaccard-Sentence**).

8.1.2. Character N-Gram Overlap

This feature explores the use of character n-grams (3-grams in this work) in predicting capturing a notion of comparability between Wikipedia documents. To extract this feature, we first processed the manually-assessed documents to perform transliteration (for Greek documents), the removal of diacritics, case folding and removal of punctuation marks. We used the Jaccard score to calculate n-gram overlap at the sentence level (F3 - **CNG-Jaccard-Sentence**). We also calculated overlap using cosine similarity based on TF-IDF (Term Frequency-Inverse Document Frequency) scores for n-grams at the document level (F4 - **CNG-TFIDF-Document**) and sentence level (F5 – **CNG-TFIDF-Sentence**).

8.1.3. Word Overlap

The sixth feature (F6 - **WordOverlap-TFIDF-Document**) computes the Jaccard score of word (cognate) overlap at document level. No translation is performed for this feature; therefore document pairs will only receive a score if they contain an exact overlap, such as shared numbers or named entities.

8.1.4. Word Length

Lastly, this feature (F7 – **WordLength-Document**) represents the size ratio with respect to the longer document. We first performed transliteration and removal of diacritics. Only words written in a Latin alphabet or numbers were counted in this process.

8.2. Evaluation Data

ACCURAT partners were asked to assess the similarity of 100 pairs of Wikipedia articles written in different languages. As a part of the exercise, assessors were asked to rate document pairs for their degree of comparability using a 5-point Likert scale (1=non-comparable; 5=parallel). Eight language pairs were evaluated in this task resulting in a total of 800 unique document pairs with judgements by two assessors for each pair. The evaluation methodology has been described in detail in D3.4 (Paramita et al. 2011).

First, we average the similarity scores given by both assessors and aggregate the average similarity scores from the Likert-scale judgements to represent two classes only: non-comparable (scores 1, 2 and 3); comparable (scores 4 and 5). The number of documents in each class is shown in Table 8 (overall language pairs) and Table 9 (pairs in each of the languages).

Table 8 Number of Documents in All Language Pairs

	Overall
Comparable Document	536 (67%)
Non-Comparable Document	264 (33%)
Total Docs	800 (100%)

Table 9 Number of Documents in Each Language Pair

	DE-EN	EL-EN	ET-EN	HR-EN	LT-EN	LV-EN	RO-EN	SL-EN
Comparable Document	64	88	41	66	67	44	70	96
Non-Comparable Document	36	12	59	34	33	56	30	4
Total Docs	100	100	100	100	100	100	100	100

As shown in the table above, the proportion of comparable and non-comparable documents are roughly similar between most language pairs, except for ET-EN and LV-EN, in which the number of non-comparable documents are higher than the comparable documents. Also, for two language pairs, the numbers of documents judged non-comparable are found to be very small: EL-EN (12%) and SL-EN (4%).

8.3. Supervised Document Classification

To evaluate our features we perform supervised document classification between pairs of Wikipedia articles compared with the human-generated judgements. The WEKA 3.6 machine learning toolkit is used to perform 10-fold cross-validation using a Naïve Bayes classifier and perform feature selection.

(1) Classification Performance

The Naïve Bayes classifier correctly classified 73.13% of the document pairs, which represents 70.71% of the comparable document pairs and 78.03% of the non-comparable documents. Confusion matrix for this classifier is shown in Table 10. Considering that all features used in this analysis were language-independent (i.e. required no external translation resources) and are easily extracted from Wikipedia articles, the results are promising.

Table 10 Confusion matrix for all language pairs

ALL PAIRS	Classified as	
	Comparable	**Non-comparable**
Comparable	379 (70.71%)	157 (29.29%)
Non-comparable	58 (21.97%)	206 (78.03%)

We also compared the classifier performance for each language pair (shown in Table 11). The first and second rows show the number of correctly classified document pairs and the total documents in that class (proportion is also shown in parentheses). The overall correctly classified instances are shown in the third row and the average F-measure in the fourth row.

As shown in the table, the classifier is able to successfully classify document pairs in different languages with an overall percentage of correctly classified instances to be above 70%. The best performance is achieved using a German-English language pair. On two language pairs (EL-EN and SL-EN), the numbers of non-comparable document pairs are low, which results in a lower classification accuracy on non-comparable documents. Similarly, on

ET-EN and LV-EN, the classification accuracy of detecting comparable documents was lower due to the small number of comparable documents in the set.

Table 11 Percentage of Correctly Classified Documents

	DE-EN	EL-EN	ET-EN	HR-EN	LT-EN	LV-EN	RO-EN	SL-EN
Comparable documents	55/64 (85.94%)	64/88 (72.73%)	25/41 (60.98%)	55/66 (83.33%)	54/67 (80.6%)	24/44 (54.55%)	55/70 (78.57%)	83/96 (86.46%)
Non-comparable documents	31/36 (86.11%)	6/12 (50%)	50/59 (84.75%)	22/34 (64.71%)	22/33 (66.67%)	47/56 (83.93%)	24/30 (80%)	1/4 (25%)
Correctly Classified Instances	**86%**	**70%**	**75%**	**77%**	**76%**	**71%**	**79%**	**84%**
	0.861	**0.747**	**0.745**	**0.769**	**0.762**	**0.702**	**0.796**	**0.88**

(2) Analysis of features

The previous section indicates that language independent features work well in classifying documents, given that there is enough training data in each class. In this section, we focus on analysing the usability of each feature in predicting comparability. First, we used `InfoGainAttributeEval`, an attribute evaluator in Weka to rank the merit of different metrics using 10-fold cross validation, and obtained the rank as shown in Table 12. Then we used a Correlation-based Feature Selection algorithm (`CfsSubsetEval`) using a `BestFirst search strategy` to evaluate different combinations of features to derive an optimal subset (feature selection). Features which were identified using this phase are shown by an asterisk (*).

To explore this further, we used each feature separately to classify the comparability level of document pairs and predict comparability. The F-measure and the number of correctly classified instances are reported in Table 12.

Table 12 Percentage of Correctly Classified Documents (sorted by the merit)

Features Name	Ranking	F-Measure**			Correctly Classified Instances**
		Comparable	Non-Comparable	Weighted Average	
WordLength-Document	1*	0.824	0.567	0.739	75%
Anchor-Jaccard-Document	2*	0.726	0.551	0.669	66%
CNG-TFIDF-Sentence	3*	0.821	0.452	0.699	73%
CNG-Jaccard-Sentence	4	0.798	0.484	0.695	71%
WordOverlap-TFIDF-Document	5	0.801	0.394	0.666	70%
CNG-TFIDF-Document	6	0.781	0.319	0.629	66.875%
Anchor-Jaccard-Sentence	7	0.767	0.231	0.59	64.25%

**Selected features by CfsSubsetEval*
***Classifier's performance when a feature is used on its own*

The data presented in the table above suggest that a combination of three features – word length, anchor overlap in document level, and character n-gram tf-idf in sentence level – are the most useful comparability metrics. When these features are used separately, they can classify comparable documents with high F-measure scores (above 0.7), however, these scores decrease on classification of non-comparable documents.

Further, we focused our evaluation on different language pairs using the same methods: `InfoGainAttributeEval` and `CfsSubsetEval`. This is aimed to analyse the merit of different metrics in different language pairs. We sorted the features by their average ranks and this is shown in Table 13.

Table 13 Ranked Features for Each Language Pair (sorted by average rank)

Features	DE-EN	EL-EN	ET-EN	HR-EN	LT-EN	LV-EN	RO-EN	SL-EN	Average Rank
CNG-Jaccard-Sentence	1*	2	4	2	1*	6	4	1	**2.63**
Anchor-Jaccard-Document	2*	4	6	4	3	2*	1*	-	**3.14**
WordLength-Document	5*	1	5*	1*	5	5	2*	-	**3.43**
CNG-TFIDF-Sentence	3*	7	1*	7	2*	1*	7	-	**4**
Anchor-Jaccard-Sentence	7	3	7	3	4	4	3	2*	**4.13**
WordOverlap-TFIDF-Document	6	5	2*	5	7	3	5	-	**4.71**
CNG-TFIDF-Document	4	6	3	6	6	7	6	-	**5.43**

The table shows that **CNG-Jaccard-Sentence** (character-n-gram extracted at the sentence level) is the most useful feature (or measure of comparability), followed by **Anchor-Jaccard-Document** (anchor overlap at the document level). Using **WordLength-Document** (ratio of document's word lengthalso proves to be a useful feature in several languages, followed by) **CNG-TFIDF-Sentence** (TF-IDF character-n-gram overlap at the sentence level). The other three features, however, have considerably lower ranks over most languages: **Anchor-Jaccard-Sentence** is less useful than **Anchor-Jaccard-Document**; **WordOverlap-TFIDF-Document** and **CNG-TFIDF-Document** also obtain low scores.

8.4. Conclusion

This section has explored a selection of language-independent features at different levels that can be used to distinguish comparable from non-comparable document pairs. Using Naïve Bayes for supervised document classification we managed to classify 70.71% of comparable text pairs correctly and 78.03% non-comparable document pairs. We can conclude that simple language-independent features, such as using the overlap of anchors (Anchor-Jaccard-Document), character n-grams (CNG-TFIDF-Sentence and CNG-Jaccard-Sentence) and word length (WordLength-Document) are suitable features for predicting whether a text pair is comparable or not. By exploring feature selection at different levels, we also conclude that anchors are best extracted at the level of the document, while scores based on character n-grams work better when extracted at the sentence level. Other features extracted at the

document level, such as character n-grams (CNG-TFIDF-Document) or word (WordOverlap-TFIDF-Document), contribute less to making classification decisions. We aim to study these features further to improve the classifier in the near future.

9. Assessing the Topical Comparability of News Corpora

In D3.4 we reported on the following work: (1) a tool for gathering comparable news texts in different language pairs; (2) an experimental method and "event relatedness scheme" for analyzing the comparability of news texts ; (3) the results of a small pilot study which tested the method and the scheme with texts in 8 ACCURAT language pairs and (4) a second, scaled up experiment, (work in progress at the time), in which we asked multiple human annotators to make judgements on 100 text pairs for 8 ACCURAT language pairs.

In this second experiment there were two aims: first, to further investigate our scheme for analyzing the comparability of news texts by asking: do people consistently agree when asked to make judgements about the different categories of news text? And second, to assess the results of our tool for gathering comparable news texts by asking: do the ranked results correlate with the categories in the scheme? We conjecture that text pairs that are about the same news event, or even more specifically about the same news event and sharing the same focal event, will contain a large amount of semantically equivalent content. Therefore, the ideal outcome would be if highly ranked news text pairs gathered by our news gathering tool are judged to be in the same news event/same focal event categories of the event relatedness scheme.

This experiment is now complete and here in D1.3, we report on the evaluation data, the results of inter-annotator agreement for the different language pairs and on the analysis of the system results

(For full details on the motivation, tools and methods for this work, please see Deliverable 3.4, Section 2, "Retrieval Techniques for General Usage Corpora", especially section 2.1, "News", and Section 2, "Evaluation", especially 2.1, "News Evaluation").

9.1. Evaluation Data

For each of 8 ACCURAT Languages (German, Croatian, Greek, Lithuanian, Latvian, Estonian, Slovenian and Romanian), we assembled 100 text pairs from deciles in the ranked list of system results (in each case the ACCURAT language was paired with English). This ensured we had example pairs for evaluation from across the full range of results.

We collected human judgements from at least 2 annotators for each of the 8 language pairs via our in-house experimental interface (see section 2.1, D 3.4).

9.2. Results: An Event Relatedness Scheme for Analysing News Texts: Agreement between Annotators

Here we report on inter-annotator agreement for the different categories in our scheme. Summary figures are shown in Table 14. Each row reports the results for two annotators' judgements over the 100 document pairs for one language. For two of the eight language pairs (Croatian and Slovenian) there were three annotators, and in these cases we report agreement results for each pair of annotators. The human judges were asked a series of questions for a set of seven questions. For each question there is a pair of columns in the table. The first column in the pair reports the percentage agreement of the two annotators on this question; the second column the raw score from which the percentage agreement was calculated, i.e. the number of document pairs for which their answer to this question was the same divided by the number of document pairs they were asked to judge.

Note that as we go across the table the denominator of the raw score goes down. This is because depending on the answers to an earlier question, a later question may not be asked. For example, if an annotator judges a document pair to be about the same news event, they will not subsequently be asked whether the document pair is about the same news event type. Furthermore, since annotators' judgement to earlier questions may diverge, only one of them may be asked a later question and in this circumstance we cannot, of course, report an agreement figure for the later questions for that document pair.

Table 14 Human Percentage Agreement on Event Relatedness Judgements

Language pair	Is News Story?		Same News Events?		Same Focal Events?		Quotes in Common?		Same News Event Type?		Same Focal Event Type?		Background in Common?	
de-en	81	81/100	89.7	70/78	75	36/48	91.7	44/48	86.4	19/22	100	22/22	90.9	20/22
el-en	88	88/100	79.5	66/83	93.5	43/46	100	46/46	80	16/20	80	16/20	60	12/20
et-en	88.5	69/78	88.3	53/60	83.3	25/30	96.7	29/30	82.6	19/23	78.3	18/23	95.7	22/23
hr-en(a1-a2)	83.7	77/92	69	49/71	70	14/20	100	20/20	82.8	24/29	37.9	11/29	75.9	22/29
hr-en(a1-a3)	90.2	83/92	89.2	66/74	65.7	23/35	94.3	33/35	67.7	21/31	93.5	29/31	77.4	24/31
hr-en(a2-a3)	82.6	76/92	67.1	47/70	66.7	14/21	85.7	18/21	73.1	19/26	30.8	8/26	76.9	20/26
lt-en	92	92/100	86.7	78/90	67.2	43/64	95.3	61/64	57.1	8/14	78.6	11/14	92.9	13/14
lv-en	92	92/100	68.9	62/90	62.5	20/32	96.9	31/32	80	24/30	73.3	22/30	90	27/30
ro-en	92.7	90/97	86.5	77/89	84.8	56/66	93.9	62/66	81.8	9/11	63.6	7/11	81.8	9/11
sl-en(a1-a2)	76	76/100	85.3	64/75	81.1	30/37	89.2	33/37	85.2	23/27	77.8	21/27	92.6	25/27
sl-en(a1-a3)	95	95/100	74.7	71/95	71.4	20/28	92.9	26/28	72.1	31/43	81.4	35/43	81.4	35/43
sl-en(a2-a3)	75	75/100	69.9	51/73	80	20/25	84	21/25	73.1	19/26	80.8	21/26	84.6	22/26
Average	86.4		79.6		75.1		93.4		76.8		73.0		83.3	

Also note that results are not given for all 100 document pairs for each language (see denominator in the "Is News Story?" raw score column). For Croatian this is because, while all of the automatically selected document pairs had passed our language identification filters, in some cases collected documents were not actually in the Croatian language. For Estonian, this is because for the time period chosen from which to gather comparable news texts, our news gathering tool simply could not find100 text pairs that exceeded its comparability threshold.

While Table 14 shows percentage agreement for each annotator pair on each question, it does not distinguish how many times the annotators agreed with a positive answer to a question as opposed to with a negative answer. This more detailed data is shown in Table 15, which shows how many times they both answered exactly positively (the "Y" columns), how many times negatively (the "N" columns) and how many times they disagreed (the "≠" columns).

9.3. Results: Assessment of the tool for Gathering Comparable News Texts

Here we investigate how the results of the tool for gathering and ranking comparable news texts correlate with different categories in our scheme for analyzing news texts.

The 100 document pairs per language were gathered using the Comparable News Retrieval Tool (CNRT) described in Deliverable 3.4. This tool aims to gather sets of document pairs in different languages on the same news story, and uses features like proximity of publication time and title similarity to achieve this aim. It provides a score for each document pair that can be interpreted as a comparability measure and used to rank the documents within topic groupings. A score threshold is used to exclude from consideration all document pairs falling below this level.

Table 15 Raw Data for Human Agreement on Event Relatedness Judgements

Language pair	Is News Story ?			Same News Events?			Same Focal Events?			Quotes in Common?			Same News Event Type?			Same Focal Event Type?			Background in Common?		
	Y	N	≠	Y	N	≠	Y	N	≠	Y	N	≠	Y	N	≠	Y	N	≠	Y	N	≠
de-en	78	3	19	48	22	8	35	1	12	1	43	4	6	13	3	0	22	0	8	12	2
el-en	83	5	12	46	20	17	43	0	3	1	45	0	8	8	4	0	16	4	2	10	8
et-en	60	9	9	30	23	7	18	7	5	3	26	1	5	14	4	3	15	5	3	19	1
hr-en (a1-a2)	71	6	15	20	29	22	14	0	6	5	15	0	18	6	5	4	7	18	11	11	7
hr-en (a1-a3)	74	9	9	35	31	8	18	5	12	8	25	2	13	8	10	0	29	2	9	15	7
hr-en (a2-a3)	70	6	16	21	26	23	13	1	7	5	13	3	16	3	7	0	8	18	9	11	6
lt-en	90	2	8	64	14	12	31	12	21	0	61	3	1	7	6	0	11	3	0	13	1
lv-en	90	2	8	32	30	28	20	0	12	2	29	1	13	11	6	0	22	0	0	27	3
ro-en	89	1	7	66	11	12	36	20	10	4	58	4	3	6	2	0	7	4	2	7	2
sl-en (a1-a2)	75	1	24	37	27	11	29	1	7	1	32	4	8	15	4	0	21	6	13	12	1
sl-en (a1-13)	95	0	5	28	43	24	19	1	8	2	24	2	8	23	12	0	35	8	20	15	8
sl-en (a2-a3)	73	2	25	25	26	22	20	0	5	1	20	4	8	11	7	0	21	5	11	11	4
Average	79.0	3.8	13.1	37.7	25.2	16.2	24.7	4.0	9.0	2.8	32.6	2.3	8.9	10.4	5.8	0.6	17.8	6.2	7.3	13.6	4.2

To investigate the correlation between CNRT's selection and ranking approach we proceeded as follows. We selected 100 document pairs per language from a variety of news story groupings, choosing 10 stories per score decile in the scoring range from the maximum comparability score down to the threshold. We used the human assessors' responses to the news event relatedness questions to place each document pair into one of four categories:

- same news event and same focal event (Same FE)
- same news event but different focal event (Same NE)
- different news event but same news event type (Same ET)
- different news event and different news event type. (Other).

We also considered a two category version of the scheme, distinguishing just same news events (regardless of whether or not the focal event is the same) from other news events (regardless of whether or not the event types are the same). We also reduced the decile level granularity in the scoring to a cruder three level division of scores ranges: those from deciles 1-3 inclusive, those from deciles 4-7, and those from deciles 8-10 (i.e. top 30%, bottom 30% and middle 40%).

The two news story category/three score range results are shown in Table 16. Several notes are in order. First, we have only considered document pairs where both annotators agree on the assignment to the news events category, i.e. where both annotators agree the documents are about the same news event or about different news events. Since they do not agree all the time, the result is differing numbers of agreed document pairs in each CNRT score decile (and hence also the 3 score range collapsed version of the decile). These differing numbers per score range mean looking at the raw numbers of document pairs in different news event categories in different score ranges is meaningless. What is significant is the ratio or proportion of same versus different news events in different score ranges. To reflect this we have normalised the numbers for each score range to show the proportion of document pairs that are about the same, as opposed to different, news events in that score range. Thus, for example, the leftmost two cells in the first row of Table 16 tell us that of the German-English document pairs returned by CNRT in the top three score deciles whose news class is agreed by both annotators, 86% are about the same news event, while 14% are about different news events.

Table 16 Distribution of Same/Different News Events by Comparability Score Range

Language Pair	Score Range 1		Score Range 2		Score Range 3		Total	
	Same	Diff	Same	Diff	Same	Diff	Same	Diff
de-en	0.86	0.14	0.63	0.37	0.58	0.42	0.69	0.31
el-en	0.5	0.5	0.86	0.14	0.64	0.36	0.68	0.32
et-en	0.56	0.44	0.55	0.45	0.67	0.33	0.57	0.43
hr-en(a1-a3)	0.65	0.35	0.65	0.35	0.18	0.82	0.53	0.47
lt-en	1	0	0.71	0.29	0.79	0.21	0.82	0.18
lv-en	0.65	0.35	0.57	0.43	0.36	0.64	0.52	0.48
ro-en	0.96	0.04	0.97	0.03	0.61	0.39	0.86	0.14
sl-en(a1-a2)	0.67	0.33	0.7	0.3	0.32	0.68	0.58	0.42
Average	0.73	0.27	0.71	0.29	0.52	0.48	0.66	0.34

The second thing to note here is that we have included only one annotator pair for each language. For the two languages (Croatian and Slovenian) where we made multiple annotator pairs, we selected that annotator pair whose average percentage agreement across all seven questions was highest.

9.4. Discussion

Annotator Agreement

Across the seven questions asked of the human assessors, agreement ranged from 73 to 93.4 percent, the average being 81%. The two questions with the lowest percentage agreement were question 3 (75.1%), which asks whether the two stories share the same focal event, and question 6 (73%), which asks whether the two stories share the same focal event type. Both

of these questions centre on the notion of focal event and low agreement suggests it may be difficult for assessors to determine what the focal event is. Highest agreement was found for question 4, which asks whether the two stories have any quotes in common. This is a relatively straightforward question to answer, so the high level of agreement is not surprising.

In order further assess the level of annotator agreement we computed the Cohen's kappa for each annotator pair and each question. The kappa scores for each question, averaged across the language pairs, range from 0 (question 6) to .6 (questions 2 and 4). While these scores would generally be interpreted as not indicating strong agreement between annotators, there are several reasons for not attaching too much weight to them. Kappa scores tend to be higher a) the more classes there are to assign observations to and b) the more equiprobable the class assignments are. In the current case there are just two classes per question, the minimum possible, and the classes are not at all equiprobable. In some cases, for example, nearly all the data is in one class about which the annotators may mostly agree; yet if the annotators disagree about the small number of examples outside the class kappa may be very low, despite high percentage agreement (in one case for question 1 we have 95% agreement and a kappa score of -0.02). In this case kappa is almost certainly unreliable as no attempt was made to select instances in both classes equally in order to fairly validate agreement – CNRT is in fact trying to select document pairs with the characteristic of being entirely in one class. Finally, in several cases the judgement sets were quite small – e.g. an average of 25 document pairs per language pair for question 6 and these small sizes render any statistic computed over them questionable.

We conclude that agreement is good – on average annotators will agree on 4 out of 5 judgements. However, if the aim is to carry out a robust assessment of inter annotator agreement for this scheme (which was not out intention here) more data better distributed over the classes would be necessary (i.e. data should be assembled to assess the scheme not by using a tool whose aim is to skew the data as far as possible). We also conclude that the notion of "focal event" may need to be further refined to enable annotators to identify it more reliably. More analysis of particular cases of disagreement and discussion with annotators would help to determine this more conclusively.

Correlation between CNRT's Comparability Score and the Event Relatedness Judgements

Table 16 shows that around 2/3 of the document pairs collected by CNRT are indeed about the same news event, while just 1/3 are not (and of the latter some will be about the same event type). Since CNRTs objective is to collect document pairs about the same news event, this is a positive result. Furthermore, note from Table 15 the high proportion of texts agreed to be about the same news event that are also about the same focal event. This can be observed by comparing the "Y" column of the "Same Focal Event" question with the "Y" column of the "Same News Event" question – the ratio ranges from just under 50% (LT) to over 93% (EL) and in all cases is substantially higher than the proportion of same news event document pairs that are not agreed to be about the same focal event. Since CNRT is designed to try to gather such same focal event texts (which we conjecture will have more comparable material), this high proportion is also satisfying and shows that CNRT is indeed doing what it is meant to do.

Looking across the score ranges we see on average two general trends. First, the percentage of document pairs which are about the same news event is highest for the top score range (deciles 1-3), slightly lower for the middle deciles (4-7) and lowest for the bottom score range (deciles 8-10). Second, concomitantly, the percentage of document pairs that are about different news event goes up as we go down the score ranges. Note that this is not true for all

language pairs, but it is true for many cases and on average. These observations suggest there is some correlation between the comparability score that CNRT uses, and news event relatedness – the higher up CNRTs ranking you got the more likely the document pairs tend to be about the same news event.

Other Observations

A few other observations are worth making at this point. The "Quotes in Common" question was designed to explore the conjecture that texts on the same news event might share quotes. Since such quotes should be genuinely parallel fragments, they would be high value elements in a comparable corpus. As we can see from the "Y" column of the "Quotes in Common Question" in Table 15, the actual numbers are not high, though as a proportion of the "Y"s of the "Same News Event" column they form around 6.5% of the text pairs on average across all language pairs which means devising techniques for extracting them could repay the effort.

When designing the event relatedness scheme, the category of "same news event type" was created because we thought it possible that retrieval tools searching for reports of a particular news event in the target language might either (a) retrieve texts in which the source language news event was mentioned in the background of another event of the same type (e.g. earthquake stories tend to mention previous earthquakes) or (b) retrieve texts whose lexis was similar to that of the source language text (so, e.g., texts about earthquakes will share many words). However, CNRT's design involves constraining its search to target language texts published very closely in time to the publication time of the source language text. It would seem unlikely, therefore, that many document pairs will be classified as "same news event type" as this requires multiple events of the same type to happen within quite a narrow time interval. If we look at Table 15 we see that for most languages fewer than 1/3 of the stories that both annotators agree are about different events are about the same news event type (compare the "N" sub-column of the "Same News Event" column with the "Yes" subcolumn of the "Same News Event Type" column). Finding any document pairs at all about the same event type, given the narrowness of the publication time window is surprising; however, Croatian and Latvian are particularly striking as for these languages even more than 1/3 of the document pairs about different events are about the same news event type. We need to examine these text pairs in detail to see why this has occurred.

More generally, while there are general trends there is considerable variability across language pairs in some of the finer detail. This is hard to interpret: it could be due to the variation in the text pairs gathered for the different languages (since the text pair collection process was driven by events in the news in the source-target language pairs during a particular time period, one would expect variation across countries); it could also be due to differences in the annotators carrying out the work and their understanding of the task (the varying scores between annotator pairs in the two languages with three annotators gives some evidence for this). More detailed examination of specific text pairs and annatotors judgements is required to investigate this further.

Implications for Further Work

Question 1 asks whether both documents are news stories (as opposed to, e.g., editorials or opinion pieces). The relatively high number of negative answers to this question (over 25% for some document sets) suggests that more work needs to be done on filtering out documents from news sources which are not news reports. Problems regarding language identification, mentioned earlier, also need to be addressed in the initial stage of assembling the comparable document sets.

Regarding the event relatedness scheme we observed above that there is less agreement between annotators in some questions than in others, particularly those relating to the notion of focal event. This issue needs further investigation to understand where the difficulty lies and perhaps the annotation guidelines will need refining,

We also observed above that the higher up CNRT's ranking you go the more likely the document pairs tend to be about the same news event. This generalization is certainly not always true and more work needs to be done to understand where and why it is not true and, consequently, how CNRT might be improved to make it even more likely to return document pairs about the same news event. One immediate impact could be to filter out those document pairs whose comparability scores are in the bottom three deciles from the comparable news corpora. Based on the evaluation results presented here, this should improve the ratio of document pairs in the ACCURAT news comparable corpora that are about the same as opposed to different news events by around 5-6%. Another idea would be to explore ways of further filtering the document pairs that CNRT returns. CNRT's strength is that it does not require full documents to be downloaded to make a decision about what news document pairs are comparable – this is a highly desirable feature computationally. However, once document pairs are deemed comparable and downloaded they could be further filtered to attempt to get the ratio of same news events to different news events higher than even the 73:27 ratio seen in the top three deciles. The document pairs assembled in the evaluation reported here could be further analyzed to gain insight into how such a filter could be constructed.

10. Conclusion

The success of extracting good-quality translation equivalences from comparable corpora to improve machine translation performance highly depends on "how comparable" the used corpora are. The task of Work Package 1 in ACCURAT aims at developing the methodology and determining criteria to measure the comparability of source and target language documents in comparable corpora. Towards this goal, we have investigated various types of information (including language-dependent, and language-independent features) which are useful in the comparability metric design (see Deliverable 1.1 for details), and presented a keyword based metric in Deliverable 1.2.

In this report, we further present two other metrics, namely machine translation based metric and lexical mapping based metric. In the machine translation based metric, the available state-of-the-art MT systems are employed for document translation, and several features are taken into account, including lexical information, document structure, keywords and named entities. In this metric, these different types of features are combined in an ensemble manner. In the lexical mapping based metric, the source language documents are translated in a word-for-word manner by using automatically generated bilingual dictionaries, and the comparability scores is computed by measuring the proportion of lexical overlapping between a source language document (now translated into target language) and a target language document with cosine similarity measure.

Using the gold standard ICC dataset created in ACCURAT for evaluation, we also validate the effectiveness of the proposed metrics. The experimental results show that both the proposed metrics can reliably predict the comparability level of comparable document pairs, given that higher comparability levels always have significantly higher comparability scores than those of lower comparability levels across different language pairs in ACCURAT. Also, due to the effect of different text translation quality between the two approaches and more information used in machine translation based metric, generally the comparability scores obtained from machine translation based metric are also significantly higher than that of lexical mapping based metric. Therefore, the results indicate that both the two metrics can be used to construct good-quality comparable corpora from the raw web crawling results. For example, filtering out less comparable document pairs in the corpora.

In addition, we also further investigate the applicability of the proposed metrics in other task. More specifically, we measure the impact of the metrics in the task of parallel phrase extraction from comparable corpora. It turns out that higher comparability scores always lead to significantly more parallel phrases extracted from comparable documents. This is also consistent with the claim that better quality of comparable corpora should have better applicability. For example, Li and Gaussier (2010) show that the performance of bilingual lexicon extraction is enhanced from the improved comparable corpora. Thus, the metrics can be applied to select more comparable documents and boost the performance of other tasks of translation equivalence extraction from comparable corpora.

Despite the encouraging results from the experiments, we also analyse the drawbacks of the proposed metrics. The main problem in machine translation based metric is that, applying MT systems for document translation is expensive as it is time-consuming, while in the keyword based metric and lexical mapping based metric, their performance highly relies on the quality of the automatically generated bilingual dictionaries. Hence, in order to overcome the existing problems to a certain degree, we also propose several ways to further improve the current metrics. This includes seeking for more linguistic resources to construct bilingual dictionaries with broader word coverage across different domains and employing

distributional semantic information of words from large-scale corpora and WordNet relations (e.g., synonym, similar-to, directly-hypernym and directly-hyponym) for lexical expansion.

Apart from the two unsupervised comparability metrics, we also present a supervised comparability metric for Wikipedia documents. The 10-fold cross validation experiments show that anchor, character n-gram and word length are the best features to predict comparability. In the evaluation of the tool (CNRT) for automatically gathering comparable news texts, encouraging results are also obtained: 2/3 of the document pairs collected by CNRT are indeed about the same news events and in general, the higher up CNRTs ranking you got the more likely the document pairs tend to be about the same news event.

Finally, given that it is an on-going project, apart from the effort in improving the performance of current metrics, in the future work we will conduct more evaluation on the proposed metric and also further explore its impact to machine translation performance. For example, we will also perform a comprehensive evaluation of the proposed metric to capture its impact on the quality of machine translation systems with phrase tables derived from comparable corpora.

11. References

Ahmet Aker et al. 2011. ACCURAT Deliverable 3.5: Tools for building comparable corpus from the Web

Bogdan Babych (a) et al. 2010. ACCURAT Deliverable 1.1: Initial report on criteria of comparability and parallelism.

Bogdan Babych (b) et al. 2010. ACCURAT Deliverable 1.2: Initial report on metrics of comparability and parallelism and their suitability.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. Bradford Books.

Voula Giouli et al. 2010. ACCURAT Deliverable 3.1: Initial Comparable Corpora.

Radu Ion et al. 2011. ACCURAT Deliverable 2.6: Toolkit for multi-level alignment and information extraction from comparable corpora.

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In proceedings of COLING 2010, Beijing, China.

Bo Li, Eric Gaussier and Akiko Aizawa. Clustering comparable corpora for bilingual lexicon extraction, In proceedings of ACL 2011, Portland, USA.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In Proceedings of COLING-ACL, 1998, Montreal, Canada.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and. Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In Proceedings of ACL 2007, Prague, Czech Republic.

Dragos Stefan Munteanu, Daniel Marcu, 2005. Improving Machine Translation Performance by Exploiting Non-parallel Corpora. In Computational Linguistics, 31(4):477-504.

Dragos Stefan Munteanu, Daniel Marcu, 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In Proceedings of COLING/ACL 2006, Sydney, Australia.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In Proceedings of ACL 2000, Hongkong, China.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment modes. In Computational Linguistics, 29(1):19-51, 2003.

Monica Paramita et al. 2011. ACCURAT Deliverable 3.4: Report on methods for collection of comparable corpora

Emmanuel Prochasson and Pascale Fung. 2011. Rare Word Translation Extraction from Aligned Comparable Documents. In Proceedings ACL-HLT 2011, Portland, USA.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German Corpora. In Proceedings of ACL 1999, Maryland, USA.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In Proceedings of ACL 1995, Cambridge, Massachusetts, USA.

Kun Yu and Junichi Tsujii. 2009. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In Proceedings of HLT-NAACL 2009, pages 121–124, Boulder, Colorado, USA.

12. List of Tables

13. List of Figures