Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents

Fangzhong Su
Centre for Translation Studies
University Of Leeds
LS2 9JT, Leeds, UK
smlfs@leeds.ac.uk

Bogdan Babych
Centre for Translation Studies
University Of Leeds
LS2 9JT, Leeds, UK
b.babych@leeds.ac.uk

Abstract

In this paper we present and evaluate three approaches to measure comparability of documents in non-parallel corpora. We develop a task-oriented definition of comparability, based on the performance of automatic extraction of translation equivalents from the documents aligned by the proposed metrics, which formalises intuitive definitions of comparability for machine translation research. We demonstrate application of our metrics for the task of automatic extraction of parallel and semi-parallel translation equivalents and discuss how these resources can be used in the frameworks of statistical and rule-based machine translation.

1 Introduction

Parallel corpora have been extensively exploited in different ways in machine translation (MT) — both in Statistical (SMT) and more recently, in Rule-Based (RBMT) architectures: in SMT aligned parallel resources are used for building translation phrase tables and calculating translation probabilities; and in RBMT, they are used for automatically building bilingual dictionaries of translation equivalents and automatically deriving bilingual mappings for frequent structural patterns. However, large parallel resources are not always available, especially for under-resourced languages or narrow domains. Therefore, in recent years, the use of cross-lingual comparable corpora has attracted considerable attention in the MT community (Sharoff et al., 2006; Fung and Cheung, 2004a; Munteanu and Marcu, 2005; Babych et al., 2008).

Most of the applications of comparable corpora focus on discovering translation equivalents to support machine translation, such as bilingual lexicon extraction (Rapp, 1995; Rapp, 1999; Morin et al., 2007; Yu and Tsujii, 2009; Li and Gaussier, 2010; Prachasson and Fung, 2011), parallel phrase extraction (Munteanu and Marcu, 2006), and parallel sentence extraction (Fung and Cheung, 2004b; Munteanu and Marcu, 2005; Munteanu et al., 2004; Smith et al., 2010).

Comparability between documents is often understood as belonging to the same subject domain, genre or text type, so this definition relies on these vague linguistic concepts. The problem with this definition then is that it cannot be exactly benchmarked, since it becomes hard to relate automated measures of comparability to such inexact and unmeasurable linguistic concepts. Research on comparable corpora needs not only good measures for comparability, but also a clearer, technologically-grounded and quantifiable definition of comparability in the first place.

In this paper we relate comparability to usefulness of comparable texts for MT. In particular, we propose a performance-based definition of comparability, as the possibility to extract parallel or quasi-parallel translation equivalents – words, phrases and sentences which are translations of each other. This definition directly relates comparability to texts' potential to improve the quality of MT by adding extracted phrases to phrase tables, training corpus or dictionaries. It also can be quantified as the rate of successful extraction of translation equivalents by automated tools, such as proposed in Munteanu and Marcu (2006).

Still, successful detection of translation equivalents from comparable corpora very much de-

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 10–19,
Avignon, France, April 23 - 27 2012.

pends 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 cross-lingual comparable documents from raw corpora crawled from 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. In this work, we directly characterize comparability by how useful comparable corpora are for the task of detecting translation equivalents in them, and ultimately to machine translation. We focus on document-level comparability, and use three categories for qualitative definition of comparability levels, defined in terms of granularity for possible alignment:

- Parallel: Traditional parallel texts that are translations of each other or approximate translations with minor variations, which can be aligned on the sentence level.

- Strongly-comparable: Texts that talk about the same event or subject, but in different languages. For example, international news about oil spill in the Gulf of Mexico, or 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: Texts in the same subject domain which describe different events. For example, customer reviews about hotel and restaurant in London. These documents do not have an independent alignment across languages, but sets of texts can be aligned on the basis of belonging to the same subject domain or sub-domain.

In this paper, we present three different approaches to measure the comparability of cross-lingual (especially under-resourced languages) comparable documents: a lexical mapping based approach, a keyword based approach, and a machine translation based approach. The experimental results show that all of them can effectively predict the comparability levels of the compared document pairs. We then further investigate the applicability of the proposed metrics by measuring their impact on the task of parallel phrase extraction from comparable corpora. It turns out that, higher comparability level predicted by the metrics consistently lead to more number of parallel phrase extracted from comparable documents. Thus, the metrics can help select more comparable document pairs to improve the performance of parallel phrase extraction.

The remainder of this paper is organized as follows. Section 2 discusses previous work. Section 3 introduces our comparability metrics. Section 4 presents the experimental results and evaluation. Section 5 describes the application of the metrics. Section 6 discusses the pros and cons of the proposed metrics, followed by conclusions and future work in Section 7.

2 Related Work

The term "comparability", which is the key concept in this work, applies to the level of corpora, documents and sub-document units. However, so far there is no widely accepted definition of comparability. For example, there is no agreement on the degree of similarity that documents in comparable corpora should have or on the criteria for measuring comparability. Also, most of the work that performs translation equivalent extraction in comparable corpora usually assumes that the corpora they use are reliably comparable and focuses on the design of efficient extraction algorithms. Therefore, there has been very little literature discussing the characteristics of comparable corpora (Maia, 2003). In this section, we introduce some representative work which tackles comparability metrics.

Some studies (Sharoff, 2007; Maia, 2003; McEnery and Xiao, 2007) analyse comparability by assessing corpus composition, such as structural criteria (e.g., format and size), and linguistic criteria (e.g., topic, domain, and genre). Kilgarriff and Rose (1998) measure similarity and homogeneity between monolingual corpora. They generate word frequency list from each corpus and then apply χ^2 statistic on the most frequent n (e.g., 500) words of the compared corpora.

The work which deals with comparability measures in cross-lingual comparable corpora is closer to our work. Saralegi et al. (2008) measure the degree of comparability of comparable corpora (English and Basque) according to the distribution of topics and publication dates of documents. They compute content similarity for all the document pairs between two corpora. These similarity scores are then input as parameters for the EMD (Earth Mover's Distance) distance measure, which is employed to calculate the global compatibility of the corpora. Munteanu and Marcu (2005; 2006) select more comparable document pairs in a cross-lingual information retrieval based manner by using a toolkit called Lemur[1]. The retrieved document pairs then serve as input for the tasks of parallel sentence and sub-sentence extraction. Smith et al. (2010) treat Wikipedia as a comparable corpus and use "interwiki" links to identify aligned comparable document pairs for the task of parallel sentence extraction. Li and Gaussier (2010) propose a comparability metric which can be applied at both document level and corpus level and use it as a measure to select more comparable texts from other external sources into the original corpora for bilingual lexicon extraction. The metric measures the proportion of words in the source language corpus translated in the target language corpus by looking up a bilingual dictionary. They evaluate the metric on the rich-resourced English-French language pair, thus good dictionary resources are available. However, this is not the case for under-resourced languages in which reliable language resources such as machine-readable bilingual dictionaries with broad word coverage or word lemmatizers might be not publicly available.

3 Comparability Metrics

To measure the comparability degree of document pairs in different languages, we need to translate the texts or map lexical items from the source language into the target languages so that we can compare them within the same language. Usually this can be done by using bilingual dictionaries (Rapp, 1999; Li and Gaussier, 2010; Prachasson and Fung, 2011) or existing machine translation tools. Based on this process, in this section we present three different approaches to measure the comparability of comparable documents.

3.1 Lexical mapping based metric

It is straightforward that we expect a bilingual dictionary can be used for lexical mapping between a language pair. However, unlike the language pairs in which both languages are rich-resourced (e.g., English-French, or English-Spanish) and dictionary resources are relatively easy to obtain, it is likely that bilingual dictionaries with good word coverage are not publicly available for under-resourced languages (e.g., English-Slovenian, or English-Lithuanian). In order to address this problem, we automatically construct dictionaries by using word alignment on large-scale parallel corpora (e.g., Europarl and JRC-Acquis[2]).

Specifically, GIZA++ toolkit (Och and Ney, 2000) with default setting is used for word alignment on the JRC-Acquis parallel corpora (Steinberger et al., 2006). The aligned word pairs together with the alignment probabilities are then converted into dictionary entries. For example, in Estonian-English language pair, the alignment example "kompanii company 0.625" in the word alignment table means the Estonian word "kompanii" can be translated as (or aligned with) the English candidate word "company" with a probability of 0.625. In the dictionary, the translation candidates are ranked by translation probability in descending order. Note that the dictionary collects inflectional form of words, but not only base form of words. This is because the dictionary is directly generated from the word alignment results and no further word lemmatization is applied.

Using the resulting dictionary, we then perform lexical mapping in a word-for-word mapping strategy. We scan each word in the source language texts to check if it occurs in the dictionary entries. If so, the first translation candidate are recorded as the corresponding mapping word. If there are more than one translation candidate, the second candidate will also be kept as the mapping result if its translation probability is higher than 0.3[3]. For non-English and English

[1] Available at http://www.lemurproject.org/

[2] The JRC-Acquis covers 22 European languages and provides large-scale parallel corpora for all the 231 language pairs.

[3] From the manual inspection on the word alignment results, we find that if the alignment probability is higher than 0.3, it is more reliable.

language pair, the non-English texts are mapped into English. If both languages are non-English (e.g., Greek-Romanian), we use English as a pivot langauge and map both the source and target language texts into English[4]. Due to the lack of reliable linguistic resources in non-English languages, mapping texts from non-English language into English can avoid language processing in non-English texts and allows us to make use of the rich resources in English for further text processing, such as stop-word filtering and word lemmatization[5]. Finally, cosine similarity measure is applied to compute the comparability strength of the compared document pairs.

3.2 Keyword based metric

The lexical mapping based metric takes all the words in the text into account for comparability measure, but if we only retain a small number of representative words (keywords) and discard all the other less informative words in each document, can we judge the comparability of a document pair by comparing these words? Our intuition is that, if two document share more keywords, they should be more comparable. To validate this, we then perform keyword extraction by using a simple TFIDF based approach, which has been shown effective for keyword or keyphrase extraction from the texts (Frank et al., 1999; Hulth, 2003; Liu et al., 2009).

More specifically, the keyword based metric can be described as below. First, similar to the lexical mapping based metric, bilingual dictionaries are used to map non-English texts into English. Thus, only the English resources are applied for stop-word filtering and word lemmatization, which are useful text preprocessing steps for keyword extraction. We then use TFIDF to measure the weight of words in the document and rank the words by their TFIDF weights in descending order. The top *n* (e.g., 30) words are extracted as keywords to represent the document. Finally, the comparability of each document pair is determined by applying cosine similarity to their keyword lists.

3.3 Machine translation based metrics

Bilingual dictionary is used for word-for-word translation in the lexical mapping based metric and words which do not occur in the dictionary will be omitted. Thus, the mapping result is like a list of isolated words and information such as word order, syntactic structure and named entities can not be preserved. Therefore, in order to improve the text translation quality, we turn to the state-of-the-art SMT systems.

In practice, we use Microsoft translation API[6] to translate texts in under-resourced languages (e.g, Lithuanian and Slovenian) into English and then explore several features for comparability metric design, which are listed as below.

- **Lexical feature**: Lemmatized bag-of-word representation of each document after stop-word filtering. Lexical similarity (denoted by W_L) of each document pair is then obtained by applying cosine measure to the lexical feature.

- **Structure feature**: 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 documents are highly comparable, their number of content words and their document length should be similar. 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 $S_{D1} <= S_{D2}$.

- **Keyword feature**: Top-20 words (ranked by TFIDF weight) of each document. keyword similarity (denoted by W_K) of two documents is also measured by cosine.

- **Named entity feature**: Named entities of each document. If more named entities co-occur in two documents, they are very likely to talk about the same event or subject and

[4]Generally in JRC-Acquis, the size of parallel corpora for most of non-English langauge pairs is much smaller than that of language pairs which contain English. Therefore, the resulting bilingual dictionaries which contain English have better word coverage as they have many more dictionary entries.

[5]We use WordNet (Fellbaum, 1998) for word lemmatization.

[6]Available at http://code.google.com/p/microsoft-translator-java-api/

thus should be more comparable. We use Stanford named entity recognizer[7] to extract named entities from the texts (Finkel et al., 2005). Again, cosine is then applied to measure the similarity of named entities (denoted by W_N) between a document pair.

We then combine these four different types of score in an ensemble manner. Specifically, a weighted average strategy is applied: each individual score is associated with a constant weight, indicating the relative confidence (importance) of the corresponding type of score. The overall comparability score (denoted by SC) of a document pair is thus computed as below:

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

where α, β, γ, and $\delta \in [0, 1]$, and $\alpha+\beta+\gamma+\delta = 1$. SC should be a value between 0 and 1, and larger SC value indicates higher comparability level.

4 Experiment and Evaluation

4.1 Data source

To investigate the reliability of the proposed comparability metrics, we perform experiments for 6 language pairs which contain under-resoured languages: German-English (DE-EN), Estonian-English (ET-EN), Lithuanian-English (LT-EN), Latvian-English (LV-EN), Slovenian-English (SL-EN) and Greek-Romanian (EL-RO). A comparable corpus is collected for each language pair. Based on the definition of comparability levels (see Section 1), human annotators fluent in both languages then manually annotated the comparability degree (parallel, strongly-comparable, and weakly-comparable) at the document level. Hence, these bilingual comparable corpora are used as gold standard for experiments. The data distribution for each language pair, i.e., number of document pairs in each comparability level, is given in Table 1.

Language pair	#document pair	parallel	strongly-comparable	weakly-comparable
DE-EN	1286	531	715	40
ET-EN	1648	182	987	479
LT-EN	1177	347	509	321
LV-EN	1252	184	558	510
SL-EN	1795	532	302	961
EL-RO	485	38	365	82

Table 1: Data distribution of gold standard corpora

4.2 Experimental results

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 gold standard comparability labels. In addition, in order to better reveal the relation between the scores obtained from the proposed metrics and comparability levels, we also measure the Pearson correlation between them[8]. For the keyword based metric, top 30 keywords are extracted from each text for experiment. For the machine translation based metric, we empirically set $\alpha = 0.5$, $\beta = \gamma = 0.2$, and $\delta = 0.1$. This is based on the assumption that, lexical feature can best characterize the comparability given the good translation quality provided by the powerful MT system, while keyword and named entity features are also better indicators of comparability than the simple document length information.

The results for the lexical mapping based metric, the keyword based metric and the machine translation based metric are listed in Table 2, 3, and 4, respectively.

Language pair	parallel	strongly-comparable	weakly-comparable	correlation
DE-EN	0.545	0.476	0.182	0.941
ET-EN	0.553	0.381	0.228	0.999
LT-EN	0.545	0.461	0.225	0.964
LV-EN	0.625	0.494	0.179	0.973
SL-EN	0.535	0.456	0.314	0.987
EL-RO	0.342	0.131	0.090	0.932

Table 2: Average comparability scores for lexical mapping based metric

Overall, from the average scores for each comparability level presented in Table 2, 3, and 4, we can see that, the scores obtained from the three comparability metrics can reli-

[7] Available at http://nlp.stanford.edu/software/CRF-NER.shtml

[8] For correlation measure, we use numerical calibration to different comparability degrees: "Parallel", "strongly-comparable" and "weakly-comparable" are converted as 3, 2, and 1, respectively. The correlation is then computed between the numerical comparability levels and the corresponding average comparability scores automatically derived from the metrics.

Language pair	parallel	strongly-comparable	weakly-comparable	correlation
DE-EN	0.526	0.486	0.084	0.941
ET-EN	0.502	0.345	0.184	0.990
LT-EN	0.485	0.420	0.202	0.954
LV-EN	0.590	0.448	0.124	0.975
SL-EN	0.551	0.505	0.292	0.937
EL-RO	0.210	0.110	0.031	0.997

Table 3: Average comparability scores for keyword based metric

Language pair	parallel	strongly-comparable	weakly-comparable	correlation
DE-EN	0.912	0.622	0.326	0.999
ET-EN	0.765	0.547	0.310	0.999
LT-EN	0.755	0.613	0.308	0.984
LV-EN	0.770	0.627	0.236	0.966
SL-EN	0.779	0.582	0.373	0.988
EL-RO	0.863	0.446	0.214	0.988

Table 4: Average comparability scores for machine translation based metric

ably reflect the comparability levels across different language pairs, as the average scores for higher comparable levels are always significantly larger than those of lower comparable levels, namely SC(parallel)>SC(strongly-comparable)>SC(weakly-comparable). In addition, in all the three metrics, the Pearson correlation scores are very high (over 0.93) across different language pairs, which indicate that there is strong correlation between the comparability scores obtained from the metrics and the corresponding comparability level.

Moreover, from the comparison of Table 2, 3, and 4, we also have several other findings. Firstly, the performance of keyword based metric (see Table 3) is comparable to the lexical mapping based metric (see Table 2) as their comparability scores for the corresponding comparability levels are similar. This means it is reasonable to determine the comparability level by only comparing a small number of keywords of the texts. Secondly, the scores obtained from the machine translation based metric (see Table 4) are significantly higher than those in both the lexical mapping based metric and the keyword based metric. Clearly, this is due to the advantages of using the state-of-the-art MT system. In comparison to the approach of using dictionary for word-for-word mapping, it can provide much better text translation which allows detecting more proportion of lexical overlapping and mining more useful features in the translated texts. Thirdly, in the lexical mapping based metric and keyword based metric, we can also see that, although the average scores for EL-RO (both under-resourced languages) conform to the comparability levels, they are much lower than those of the other 5 language pairs. The reason is that, the size of the parallel corpora in JRC-Acquis for these 5 language pairs are significantly larger (over 1 million parallel sentences) than that of EL-EN, RO-EN[9], and EL-RO, thus the resulting dictionaries of these 5 language pairs also contain many more dictionary entries.

5 Application

The experiments in Section 4 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 equivalent detection from comparable corpora? In this section, we further measure the impact of the metrics on parallel phrase extraction (**PPE**) from comparable corpora. Our intuition is that, if document pairs are assigned higher comparability scores by the metrics, they should be more comparable and thus more parallel phrases can be extracted from them.

The algorithm of parallel phrase extraction, which develops the approached presented in Munteanu and Marcu (2006), uses lexical overlap and structural matching measures (Ion, 2012). Taking a list of bilingual comparable document pairs as input, the extraction algorithm involves the following steps.

1. Split the source and target language documents into phrases.

2. Compute the degree of parallelism for each candidate pair of phrases by using the bilingual dictionary generated from GIZA++ (base dictionary), and retain all the phrase pairs with a score larger than a predefined parallelism threshold.

[9]Remember that in our experiment, English is used as the pivot language for non-English langauge pairs.

3. Apply GIZA++ to the retained phrase pairs to detect new dictionary entries and add them to the base dictionary.

4. Repeat Step 2 and 3 for several times (empirically set at 5) by using the augmented dictionary, and output the detected phrase pairs.

Phrases which are extracted by this algorithm are frequently not exact translation equivalents. Below we give some English-German examples of extracted equivalents with their corresponding alignment scores:

1. But a successful mission — seiner überaus erfolgreichen Mission abgebremst — 0.815501989333333

2. Former President Jimmy Carter — Der ehemalige US-Präsident Jimmy Carter — 0.69708324976825

3. on the Korean Peninsula — auf der koreanischen Halbinsel — 0.8677432145

4. across the Muslim world — mit der muslimischen Welt ermöglichen — 0.893330864

5. to join the United Nations — der Weg in die Vereinten Nationen offensteht — 0.397418711927629

Even though some of the extracted phrases are not exact translation equivalents, they may still be useful resources both for SMT and RBMT if these phrases are passed through an extra preprocessing stage, of if the engines are modified specifically to work with semi-parallel translation equivalents extracted from comparable texts. We address this issue in the discussion section (see Section 6).

For evaluation, we measure how the metrics affect the performance of parallel phrase extraction algorithm on 5 language pairs (DE-EN, ET-EN, LT-EN, LV-EN, and SL-EN). A large raw comparable corpus for each language pair was crawled from the Web, and the metrics were then applied to assign comparability scores to all the document pairs in each corpus. For each language pair, we set three different intervals based on the comparability score (SC) and randomly select 500 document pairs in each interval for evaluation. 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. For the lexical mapping based metric and keyword based metric, since their scores are lower than those of the MT based metric for each comparability level, we set 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[10], and computes the average number of extracted phrases per 100000 words (the sum of words in the source and target language documents) for each interval. In addition, the Pearson correlation measure is also applied to measure the correlation between the interval[11] of comparability scores and the number of extracted parallel phrases. The results which summarize the impact of the three metrics to the performance of parallel phrase extraction are listed in Table 5, 6, and 7, respectively.

Language pair	0.1<= SC<0.2	0.2<= SC<0.4	SC>=0.4	correlation
DE-EN	728	1434	2510	0.993
ET-EN	313	631	1166	0.989
LT-EN	258	419	894	0.962
LV-EN	470	859	1900	0.967
SL-EN	393	946	2220	0.975

Table 5: Impact of the lexical mapping based metric to parallel phrase extraction

Language pair	0.1<= SC<0.2	0.2<= SC<0.4	SC>=0.4	correlation
DE-EN	1007	1340	2151	0.972
ET-EN	438	650	1050	0.984
LT-EN	306	442	765	0.973
LV-EN	600	966	1722	0.980
SL-EN	715	1026	1854	0.967

Table 6: Impact of the keyword based metric to parallel phrase extraction

From Table 5, 6, and 7, we can see that for all the 5 language pairs, based on the average number of extracted aligned phrases, clearly we have interval (3)>(2)>(1). In other words, in any of the three metrics, a higher comparability level always leads to significantly more number

[10] A manual evaluation of a small set of extracted data shows that parallel phrases with parallelism score >=0.4 are more reliable.

[11] For the purpose of correlation measure, the three intervals are numerically calibrated as "1", "2", and "3", respectively.

Language pair	0.1<= SC<0.3	0.3<= SC<0.5	SC>=0.5	correlation
DE-EN	861	1547	2552	0.996
ET-EN	448	883	1251	0.999
LT-EN	293	483	1070	0.959
LV-EN	589	1072	2037	0.982
SL-EN	560	1151	2421	0.979

Table 7: Impact of the machine translation based metric to parallel phrase extraction

of aligned phrases extracted from the comparable documents. Moreover, although the lexical mapping based metric and the keyword based metric produce lower comparability scores than the MT based metric (see Section 4), they have similar impact to the task of parallel phrase extraction. This means, the comparability score itself does not matter much, as long as the metrics are reliable and proper thresholds are set for different metrics.

In all the three metrics, the Pearson correlation scores are very close to 1 for all the language pairs, which indicate that the intervals of comparability scores obtained from the metrics are in line with the performance of equivalent extraction algorithm. Therefore, in order to extract more parallel phrases (or other translation equivalents) from comparable corpora, we can try to improve the corpus comparability by applying the comparability metrics beforehand to add highly comparable document pairs in the corpora.

6 Discussion

We have presented three different approaches to measure comparability at the document level. In this section, we will analyze the advantages and limitations of the proposed metrics, and the feasibility of using semi-parallel equivalents in MT.

6.1 Pros and cons of the metrics

Using bilingual dictionary for lexical mapping is simple and fast. However, as it adopts the word-for-word mapping strategy and out-of-vocabulary (OOV) words are omitted, the linguistic structure of the original texts is badly hurt after mapping. Thus, apart from lexical information, it is difficult to explore more useful features for the comparability metrics. The TFIDF based keyword extraction approach allows us to select more representative words and prune a large amount of less informative words from the texts. The keywords are usually relevant to subject and domain terms, which is quite useful in judging the comparability of two documents. Both the lexical mapping based approach and the keyword based approach use dictionary for lexical translation, thus rely on the availability and completeness of the dictionary resources or large scale parallel corpora.

For the machine translation based metric, it provides much better text translation than the dictionary-based approach so that the comparability of two document can be better revealed from the richer lexical information and other useful features, such as named entities. However, the text translation process is expensive, as it depends on the availability of the powerful MT systems[12] and takes much longer than the simple dictionary based translation.

In addition, we use a translation strategy of translating texts from under-resourced (or less-resourced) languages into rich-resourced language. In case that both languages are under-resourced languages, English is used as the pivot langauge for translation. This can compensate the shortage of the linguistic resources in the under-resourced languages and take advantages of various resources in the rich-resourced languages.

6.2 Using semi-parallel equivalents in MT systems

We note that modern SMT and RBMT systems take maximal advantage of strictly parallel phrases, but they still do not use full potential of the semi-parallel translation equivalents, of the type that is illustrated in the application section (see Section 5). Such resources, even though they are not exact equivalents contain useful information which is not used by the systems.

In particular, the modern decoders do not work with under-specified phrases in phrase tables, and do not work with factored semantic features. For example, the phrase:

But a successful mission — seiner überaus erfolgreichen Mission abgebremst

The English side contains the word *but*, which pre-supposes contrast, and on the Greman side words *überaus erfolgreichen* (“generally successful”) and *abgebremst* (“slowed down”) – which taken together exemplify a contrast, since they

[12] Alternatively, we can also train MT systems for text translation by using the available SMT toolkits (e.g., Moses) on large scale parallel corpora.

have different semantic prosodies. In this example the semantic feature of contrast can be extracted and reused in other contexts. However, this would require the development of a new generation of decoders or rule-based systems which can successfully identify and reuse such subtle semantic features.

7 Conclusion and Future work

The success of extracting good-quality translation equivalents from comparable corpora to improve machine translation performance highly depends on "how comparable" the used corpora are. In this paper, we propose three different comparability measures at the document level. The experiments show that all the three approaches can effectively determine the comparability levels of comparable document pairs. We also further investigate the impact of the metrics on 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. Since better quality of comparable corpora should have better applicability, our metrics can be applied to select highly comparable document pairs for the tasks of translation equivalent extraction.

In the future work, we will conduct more comprehensive evaluation of the metrics by capturing its impact to the performance of machine translation systems with extended phrase tables derived from comparable corpora.

Acknowledgments

We thank Radu Ion at RACAI for providing us the toolkit of parallel phrase extraction, and the three anonymous reviewers for valuable comments. This work is supported by the EU funded ACCURAT project (FP7-ICT-2009-4-248347) at the Centre for Translation Studies, University of Leeds.

References

Bogdan Babych, Serge Sharoff and Anthony Hartley. 2008. *Generalising Lexical Translation Strategies for MT Using Comparable Corpora*. Proceedings of LREC 2008, Marrakech, Morocco.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. *Looking for candidate translational equivalents in specialized, comparable corpora*. Proceedings of COLING 2002, Taipei, Taiwan.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Jenny Finkel, Trond Grenager, and Christopher Manning. 2005. *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*. Proceedings of ACL 2005, University of Michigan, Ann Arbor, USA.

Eibe Frank, Gordon Paynter and Ian Witten. 1999. *Domain-specific keyphrase extraction*. Proceedings of IJCAI 1999, Stockholm, Sweden.

Pascale Fung and Percy Cheung. 2004a. *Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM*. Proceedings of EMNLP 2004, Barcelona, Spain.

Pascale Fung and Percy Cheung. 2004b. *Multi-level bootstrapping for extracting parallel sentences from a quasicomparable corpus*. Proceedings of COLING 2004, Geneva, Switzerland.

Anette Hulth. 2003. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge*. Proceedings of EMNLP 2003, Sapporo, Japan.

Radu Ion. 2012. *PEXACC: A Parallel Data Mining Algorithm from Comparable Corpora*. Proceedings of LREC 2012, Istanbul, Turkey.

Adam Kilgarriff and Tony Rose. 1998. *Measures for corpus similarity and homogeneity*. Proceedings of EMNLP 1998, Granada, Spain.

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

Feifan Liu, Deana Pennell, Fei Liu and Yang Liu. 2009. *Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts*. Proceedings of NAACL 2009, Boulder, Colorado, USA.

Belinda Maia. 2003. *What are comparable corpora?* Proceedings of the Corpus Linguistics workshop on Multilingual Corpora: Linguistic requirements and technical perspectives, 2003, Lancaster, U.K.

Anthony McEnery and Zhonghua Xiao. 2007. *Parallel and comparable corpora?* In Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters, Clevedon, UK.

Emmanuel Morin, Beatrice Daille, Korchi Takeuchi and Kyo Kageura. 2007. *Bilingual terminology mining — using brain, not brawn comparable corpora*. Proceedings of ACL 2007, Prague, Czech Republic.

Dragos Munteanu and Daniel Marcu. 2006. *Extracting parallel sub-sentential fragments from non-parallel corpora*. Proceedings of ACL 2006, Syndey, Australia.

Dragos Munteanu and Daniel Marcu. 2005. *Improving machine translation performance by exploiting non-parallel corpora*. Computational Linguistics, 31(4): 477-504.

Dragos Munteanu, Alexander Fraser and Daniel Marcu. 2004. *Improved machine translation performance via parallel sentence extraction from comparable corpora*. Proceedings of HLT-NAACL 2004, Boston, USA.

Franz Och and Hermann Ney. 2000. *Improved Statistical Alignment Models*. Proceedings of ACL 2000, Hongkong, China.

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

Reinhard Rapp. 1995. *Identifying Word Translation in Non-Parallel Texts*. Proceedings of ACL 1995, Cambridge, Massachusetts, USA.

Reinhard Rapp. 1999. *Automatic identification of word translations from unrelated English and German corpora*. Proceedings of ACL 1999, College Park, Maryland, USA.

Xabier Saralegi, Inaki Vicente and Antton Gurrutxaga. 2008. *Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain*. Proceedings of the Workshop on Comparable Corpora, LREC 2008, Marrakech, Morocco.

Serge Sharoff. 2007. *Classifying Web corpora into domain and genre using automatic feature identification*. Proceedings of 3rd Web as Corpus Workshop, Louvain-la-Neuve, Belgium.

Serge Sharoff, Bogdan Babych and Anthony Hartley. 2006. *Using Comparable Corpora to Solve Problems Difficult for Human Translators*. Proceedings of ACL 2006, Syndey, Australia.

Jason Smith, Chris Quirk and Kristina Toutanova. 2010. *Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment*. Proceedings of NAACL 2010, Los Angeles, USA.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat and Dan Tufis. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of LREC 2006, Genoa, Italy.

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