Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles

Monica Lestari Paramita, Paul Clough, Ahmet Aker and Robert Gaizauskas
University of Sheffield
211 Portobello Street, Sheffield, United Kingdom
m.paramita@sheffield.ac.uk, p.d.clough@sheffield.ac.uk, a.aker@dcs.shef.ac.uk, r.gaizauskas@sheffield.ac.uk

Abstract

Wikipedia articles in different languages have been mined to support various tasks, such as Cross-Language Information Retrieval (CLIR) and Statistical Machine Translation (SMT). Articles on the same topic in different languages are often connected by inter-language links, which can be used to identify similar or comparable content. In this work, we investigate the correlation between similarity measures utilising language-independent and language-dependent features and respective human judgments. A collection of 800 Wikipedia pairs from 8 different language pairs were collected and judged for similarity by two assessors. We report the development of this corpus and inter-assessor agreement between judges across the languages. Results show that similarity measured using language independent features is comparable to using an approach based on translating non-English documents. In both cases the correlation with human judgments is low but also dependent upon the language pair. The results and corpus generated from this work also provide insights into the measurement of cross-language similarity.

Keywords: Wikipedia, cross-language similarity, evaluation

1. Introduction

Wikipedia has been mined for various linguistic purposes because of the diversity and richness of information available in a variety of languages (Tomás et al., 2010). In addition, the presence of inter-language links, which connect documents from different languages describing the same topic, makes Wikipedia a useful multilingual resource (e.g. as a source of comparable documents). However, although articles written in different languages on the same topic could be considered comparable (Gamallo & López, 2010), the degree of similarity may vary widely. Parts of the content could be translation equivalents (i.e. parallel); other parts may have been developed independently and share little thematic or lexical overlap. For tasks, such as Cross-Language Information Retrieval (CLIR) or Statistical MT (SMT), the degree of similarity between texts will affect the quality of translation resources subsequently created; using non-similar documents will introduce noise and reduce MT performance (Lu et al., 2007).

Different measures have been developed to measure the similarity between Wikipedia articles in different languages (see Section 2) which can be used to filter out non-similar documents. However, little past work has analysed whether or not these methods correlate with human assessments across multiple languages. In this work we have collected manual judgments on Wikipedia articles in various language pairs, which include 7 under-resourced languages. We analyse the judgments gathered for inter-assessor agreement and compare the judgments with two measures of document-level similarity based on using language dependent and language-independent features. Being able to reliably measure the similarity of Wikipedia articles across languages would assist in using Wikipedia as a source of comparable data.

This paper is structured as follows: Section 2 provides a summary of past work on comparing the similarity between Wikipedia articles; Section 3 describes the methodology used in our experiments including the creation of human cross-language similarity judgments; Section 4 discusses results obtained from comparing Wikipedia articles across languages; Section 5 provides a discussion of results; finally Section 6 concludes the paper and provides directions for further research.

2. Previous Work

Wikipedia is often viewed as a promising source of comparable documents as pairings of similar (or near similar) documents in different languages are provided through the inter-language links (Otero & López, 2010). However, Wikipedia articles on the same topic are not necessarily equivalent to each other and in some cases the entry description may even contain information which is contradictory (Filatova, 2009). Nevertheless, Wikipedia does contain a rich amount of information that can be mined. For example, titles from articles connected by inter-language links have been extracted and used as the source of bilingual lexicons, enabling parallel sentences within connected articles to be identified without the use of any other linguistic resources (Adafre & de Rijke, 2006; Erdmann et al., 2008; Tomás et al., 2010). Smith et al. (2010) also used Wikipedia as a source of similar or comparable sentences but instead used the image captions. Lin et al. (2011) mined information found in the infoboxes to gather named entities and other information in different languages.

Adafre & de Rijke (2006) developed a method to retrieve parallel sentences from Wikipedia documents by using information about the overlap of anchors. Smith et al. (2010) developed this idea by using additional features, such as sentence length and longest aligned/unaligned words to develop a binary classifier trained on parallel corpora. Bharadwaj & Varma (2011) also developed a binary sentence classifier for English-Hindi which does not require parallel corpora or other linguistic resources. They first indexed the content

of documents, treating each sentence as a bag-of-words and creating separate indexes for each language. To identify whether a sentence pair was parallel or not, they performed retrieval for each sentence from the appropriate index, i.e. English sentences queried on the English index; Hindi sentences queried on the Hindi index. Different features were then extracted, such as the intersection and union of retrieved articles and sentence lengths. They report that the binary sentence classifier is able to identify parallel sentences with an accuracy of 78%.

Several methods have also been used to assess the accuracy of extracted information from Wikipedia. For example, Yu & Tsujii (2009) conducted human evaluation to assess the accuracy of extracted parallel phrases; whilst Smith et al. (2010) and Adafre & de Rijke (2006) conducted similar evaluations at the sentence level. Comparable corpora are mostly evaluated by calculating the improvement of MT performance (Munteanu & Marcu, 2005).

However, despite the continued interest in Wikipedia there seems to be little work on comparing similarity at the document level in Wikipedia. One paper that does consider document-level similarity attempts to identify parallel documents from Wikipedia (Patry & Langlais, 2011). The method first retrieves candidate document pairs using an Information Retrieval system. Parallel documents are identified using lightweight content-based features extracted from the documents, such as numbers, words only occurring once (hapax legomena) and punctuation marks. They report that the resulting classifier can correctly identify parallel and noisy parallel documents with an accuracy of 80%.

Much of the previous work has been conducted based on the English Wikipedia. However, given the variance in size and interconnectivity of Wikipedia in different languages, the performance of similarity measures is likely to vary (particularly for languages where there exist limited translation resources). This paper aims to address this and provide empirical evidence demonstrating the success of measuring cross-language similarity between different language pairs. In addition, to the best of our knowledge, there has been little or no research on comparing automatically-derived similarity scores and human judgments.

3. Methodology

3.1 Document Pre-Processing

Articles from dumps of Wikipedia[1] were downloaded for 7 under-resourced language pairs[2] and articles linked through inter-language links were extracted using JWPL (Zeesh et al., 2008). In these experiments we have used the following language pairs: Croatian-English (HR-EN), Estonian-English (ET-EN), Greek-English (EL-EN), Latvian-English (LV-EN), Lithuanian-English (LT-EN), Romanian-English (RO-EN), and Slovenian-English (SL-EN). All of these languages have limited translation resources available and would benefit from language-independent methods of assessing cross-language similarity. We also included one additional pair, German-English (DE-EN), to compare performance against as a language pair that is well-resourced and for which high-quality translation resources are available.

Wikipedia articles were pre-processed with information, such as infoboxes, images, tables, etc., filtered out. Plaintext only from the main body of Wikipedia articles was extracted and used as the basis for human cross-language similarity judgments.

Lang	Number of documents		Number of entries in bilingual lexicon
	Total	Linked to EN	
DE	1,036,144	637,382	181,408
EL	49,275	36,752	28,294
ET	72,231	42,008	22,645
HR	81,366	51,432	26,804
LT	102,407	57,954	41,497
LV	26,297	21,302	15,511
RO	141,284	97,815	35,774
SL	85,709	51,332	25,101

Table 1: Size of initial Wikipedia datasets

Table 1 shows the statistics of Wikipedia dumps used in this study. The second column shows the total number of articles in each language. The third column shows for each language the number of articles that are linked to an English article on the same topic using inter-language links. The last column shows the number of entries in the bilingual lexicon used in the similarity measures described in Section 3.2.

3.2 Similarity Measures

Two approaches for assessing document-level similarity between Wikipedia articles written in different languages were investigated: a language-independent approach based on using a bilingual lexicon derived from Wikipedia (referred to as *Anchor+word overlap*); a second approach that involved translating all non-English documents into English using available MT systems[3] (referred to as *Translation*). The latter approach enabled comparison with machine translation, however in practice is not viable due to the limited availability of translation resources.

The first approach, similar to Adafre & de Rijke (2006), determines sentence similarity by measuring overlap of anchor texts and cognates (e.g. numbers, dates and named entities) which appear as the same text string in different language versions of the text (see example in Figure 1). To translate the anchors, we start by extracting all document titles (typically nouns, named entities or

[1]Data downloaded March 2010: http://dumps.wikimedia.org/

[2]Providing translation resources for under-resourced languages is the goal of the ACCURAT (http://www.accurat-project.eu/) project within which this study was carried out.

[3]Bing Translate was used to translate all document pairs apart from HR-EN, which was translated using Google Translate.

phrases) which are connected using inter-language links and using them to build a bilingual title lexicon for each language pair (e.g. '*asteroidov*' ⇔ '*asteroid*' for Slovenian and English). We then use the lexicon to translate all anchor texts in the non-English Wikipedia article into English. We measure the proportion of overlapping terms using Jaccard coefficient; each sentence is treated as binary vectors (or sets) such that only token types are counted. Figure 1 shows an original non-English article (in Slovenian) where anchor texts are shown in bold[4]. Using the bilingual lexicon the anchor texts are replaced with their English equivalent. The Slovenian text is then compared for the overlap of terms with the equivalent English article where cognates (e.g. numbers in Figure 1) are also compared. The second approach also measures overlap of terms in sentence level but instead of using anchor+word overlap, it measures term overlap between the original English text and the English translation of the non-English text.

Original Slovenian text (anchor texts in bold)
Večinajih je v bližini**[[družinaVesta|asteroidnedružineVesta]]**. Imajopodobne**[[izsrednost|izsrednosti]]**, todanjihova**[[elipsa|velikapolos]]**leži v območju od 2,18**[[astronomskaenota|a. .e.]]** do 2,50 a. e. (kjer je **[[Kirkwoodovavrzel|Kirkwoodovavrzel]]** 3 : 1).

Slovenian text with anchors replaced with English (bold)
Večinajih je v bližini**[[vesta family]]**. Imajopodobne**[[eccentricity]]**, todanjihova**[[ellipsis]]**leži v območju od 2,18**[[astronomical unit]]** do 2,50 a. e. (kjer je **[[kirkwood gap]]** 3 : 1)

Equivalent English article (matches in bold)
A large proportion have orbital elements similar to those of 4 Vesta, either close enough to be part of the **[[vesta family]]**, or having similar **[[eccentricity (orbit)]]** and **[[inclination]]**s but with a **[[semi-major axis]]** lying between about **2.18[[astronomical unit]]** and the **3:1[[kirkwood gap]]** at **2.50** AU.

Figure 1: Example anchor text translation

In both approaches we perform a pairwise comparison between all sentence pairs allowing a *1:1* and *M:1* correspondence between sentences in both articles. We first split documents into sentences and for each sentence in the shorter Wikipedia article, we calculate its similarity with all sentences in the longer document. The sentence is paired with the sentence receiving the highest similarity score. Different sentences in the shorter article may be paired to the same sentence in the longer document. This accommodates cases in which simpler sentences are combined into an equivalent complex sentence (an example of this is also shown in Figure 1). A minimum similarity threshold is set below which sentence pairs are ignored. This threshold was set based on manual inspection of aligned sentence pairs. The local similarity scores (the similarity scores between sentence pairs) are combined into a global (or document-level) score by computing the mean value of all aligned sentence pair scores normalised by the number of sentences in the shorter document.

3.3 Eliciting Human Judgments

To select articles for human inspection, we first sorted all Wikipedia articles for a given language pair by their overall anchor and word overlap similarity score (Section 3.2). The scores were divided into 10 bins and we randomly selected 10 document pairs from each bin[5] resulting in a total of 100 documents per language pair. This initial selection process was undertaken to provide a range of article pairs with different similarity scores to include in the test data. In total 97% of articles included in the dataset contain fewer than 1,000 tokens to ensure judges were able to read and digest the articles in a reasonable time and to limit assessor fatigue. Table 2 provides a summary of the documents used for human similarity judgement.

Total number of documents	1,600 (800 pairs)
Number of languages	9 (DE, EL, EN, ET, HR, LT, LV, RO & SL)
Average number of words per document	450.59 (min: 107, max: 1,546)
Average number of sentences per document	51.31 (min: 22, max: 1,028)

Table 2: Summary of documents used for human similarity judgments

Given a pair of Wikipedia articles in a language pair, we asked assessors to read the articles and answer the four questions[6] shown in Figure 2. Assessors (16 in total) were all native speakers of the non-English language and fluent speakers of English. We used 5-point Likert scales for all questions. For the questions regarding an assessment of document-level similarity (and comparability) we did not provide descriptions for each level. A general definition of similarity is complex (Hatzivassiloglou et al., 1999); therefore by using a scale and asking assessors to comment on characteristics they felt contributed to their judgment of similarity (see Q1) we can better understand what characterises cross-language similarity between Wikipedia articles (see Section 5). All judges were partners in the ACCURAT project and therefore have a reasonable degree of knowledge about comparability and similarity. The documents and human judgments are available for public download[7].

[4]Note: the text in bold that appears with a '|' character separating terms represents the referred article title and the document text as it appears to the user.

[5]When this was not possible (i.e. fewer than 10 document pairs were found in a bin), the maximum number of document pairs in that bin were chosen for the evaluation set and a higher number of documents were chosen from the lower bins to achieve the total number of 100 document pairs.

[6]The questions were based a prior pilot study in which 10 assessors assessed 5 document pairs and gave comments on the evaluation scheme and decisions regarding their assigned similarity score.

[7]Data and judgments are available for download from here: http://ir.shef.ac.uk/cloughie/resources/similarity_corpus.html

Q1. How similar are these two documents?

○ 1 (very different) ○ 2 ○ 3 ○ 4 ○ 5 (very similar)

Why did you give this similarity score (please tick all relevant ones):

☑ Documents contain similar structure or main sections
☐ Documents contain overlapping named entities
☐ Fragments (e.g. sentences) of one document can be aligned to the other
☐ Content in one document seems to be derived or translated from the other
☐ Documents contain different information (e.g. different perspective, aspects, areas)
☐ Others, please mention: .....

Q2. What proportion of overall document contents is shared between the documents?

○ 1 (none) ○ 2 ○ 3 ○ 4 ○ 5 (all)

Q3. Of the shared content (if there is any), on average how similar are the matching sentences?

○ 1 (very different) ○ 2 ○ 3 ○ 4 ○ 5 (very similar)

Q4. Overall, what is the comparability level between these two documents?

○ 1 (very different) ○ 2 ○ 3 ○ 4 ○ 5 (very similar)

Figure 2: Evaluation sheet completed by human assessors for each document pair

4. Results and Analysis

4.1 Responses to the Questionnaire

There is a significant correlation between the similarity level (Q1) assigned by the assessor and the level of comparability (Q4) (*rho=0.873; p<0.01*) and the similarity level (Q1) and the overall proportion of shared content (Q2) (*rho=0.900; p<0.01*) suggesting that the more overlap between information in article pairs the greater the perceived degree of similarity. There is also a significant correlation between the overall similarity score (Q1) and the similarity between matching sentences of the shared content (Q3) (*rho=0.727; p<0.01*).

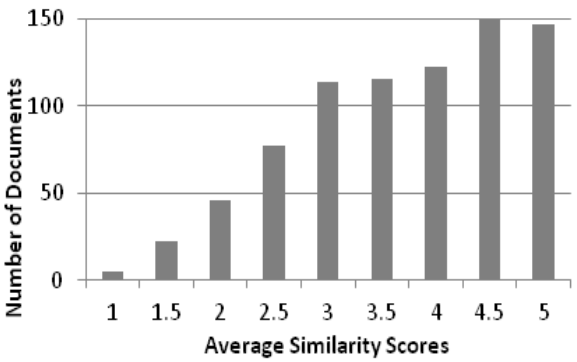

Figure 3: Distribution of document-level similarity scores averaged across both assessors (*N=800*)

Figure 3 shows the distribution of the similarity scores assigned to document pairs where multiple scores are averaged across the scores of the two assessors. For the 800 document pairs, 52.5% of document pairs are judged to exhibit a high degree of similarity (average score equals 4 or above), 28.8% judged to be moderately similar (average score between 2.5 and 3.5) and 18.8% judged to be different (average score of 2 and less). This confirms that articles in different languages on the same topic are not necessarily similar and therefore a suitable method to identify cross-language similarity is required. We explore in more detail what features judges use to derive their judgment of similarity in Section 5.

4.2 Inter-Assessor Agreement

We report the agreement between assessors for each question in the evaluation task (shown in Figure 2). Scores are calculated over the original 5-point scale[8] and also for a 2-point scale created by aggregating the results for scores 1-3 (low similarity) and 4-5 (high similarity). The values in parentheses represent the proportion of cases where assessors' scores are the same.

Question	Weighted Cohen's Kappa (5 classes)	Cohen's Kappa (2 classes)
Q1) Similarity	0.38 (41%)	0.43 (73%)
Q2) Proportion	0.47 (48%)	0.52 (77%)
Q3) Similar Sentences	0.39 (50%)	0.42 (81%)
Q4) Comparability Level	0.37 (48%)	0.46 (80%)

Table 3: Inter-assessor agreement (% indicates the proportion of times assessors agree on the same value)

As shown in Table 3, assessors chose the same similarity score to represent document pairs 41% of the time. However upon further inspection we find that in

[8] Agreement for the 5 similarity levels is calculated using a weighted version of Cohen's Kappa, in which the order of classes is taken into account, e.g. similarity scores of 1 and 2 are in better agreement than scores 1 and 5.

cases when assessors disagree, 44% of scores assigned to the document pairs differ only by 1 (14% by 2, 2% by 3 and 0.4% by 4). This is shown by the high level of agreement when considering scores combined into 2 classes. The proportion of cases in which assessors give the same value rises to 73%.

Table 4 shows agreement between assessors for each of the questions on a 5-point scale broken down by language pairs. There is considerable variance across language pairs from 25% agreement (DE-EN) to 70% (SL-EN) for assigned similarity scores. Table 5 shows inter-assessor agreement when scores are combined into 2 classes. In all cases the agreement is markedly improved.

Lang	Similarity	Proportion	Similar Sentences	Comparability Level
DE-EN	0.34 (25%)	0.45 (46%)	0.45 (52%)	0.33 (42%)
ET-EN	0.49 (57%)	0.49 (58%)	0.36 (45%)	0.44 (69%)
EL-EN	0.25 (43%)	0.36 (50%)	0.37 (56%)	0.41 (59%)
HR-EN	0.34 (28%)	0.38 (34%)	0.43 (51%)	0.25 (24%)
LT-EN	0.14 (19%)	0.31 (43%)	0.14 (27%)	0.08 (23%)
LV-EN	0.43 (45%)	0.36 (39%)	0.45 (51%)	0.31 (43%)
RO-EN	0.37 (37%)	0.33 (38%)	0.39 (48%)	0.52 (59%)
SL-EN	0.36 (70%)	0.62 (79%)	0.20 (72%)	0.30 (65%)

Table 4: Inter-assessor agreement (weighted Cohen's Kappa) for each language pair (5 classes)

Lang	Similarity	Proportion	Similar Sentences	Comparability Level
DE-EN	0.58 (79%)	0.6 (80%)	0.44 (80%)	0.12 (70%)
ET-EN	0.71 (86%)	0.67 (84%)	0.42 (75%)	0.49 (98%)
EL-EN	0.14 (64%)	0.21 (64%)	0.27 (75%)	0.68 (88%)
HR-EN	0.22 (53%)	0.45 (70%)	0.35 (82%)	0.42 (82%)
LT-EN	0.16 (53%)	0.43 (71%)	0.35 (75%)	0.00 (61%)
LV-EN	0.55 (78%)	0.53 (78%)	0.49 (81%)	0.27 (83%)
RO-EN	0.39 (72%)	0.34 (69%)	0.57 (84%)	0.58 (85%)
SL-EN	0.71 (97%)	0.75 (97%)	0.21 (93%)	0.31 (71%)

Table 5: Inter-assessor agreement (Cohen's Kappa) for each language pair (2 classes)

4.3 Correlation of Similarity Measures to Human Judgments

Section 3.2 described two approaches to compute cross-language similarity between document pairs: the first a language-independent approach; the second based on translation of non-English articles into English and computing monolingual similarity. Similarity values between the two approaches are highly correlated (*rho=0.744*, *p<0.01*) showing that results obtained using language-independent features are comparable to results based on having the availability of translation resources. A scatter plot of the scores obtained from these two approaches for all document pairs is shown in Figure 4.

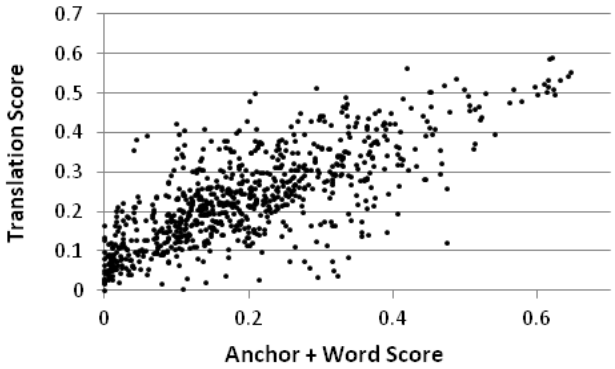

Figure 4: Correlation between anchor+word overlap and similarity based on document translation

Table 6 shows the correlation between similarity measures and sets of judgments: those from each assessor separately and combined (the mean score of each judgment rounded up to the nearest whole number). In these results we use human judgments based on 5-point Likert scale. From the results for the combined judgments the anchor and word overlap approach shows a higher correlation (*rho=0.353, p<0.01*) than the approach based on translating non-English articles into English and computing word overlap (*rho=0.325, p<0.01*).

Judgment Set	Anchor+word overlap	Translation
Judgment 1	0.290	0.228
Judgment 2	0.321	0.323
Combined	0.353	0.325

Table 6: Correlation (Spearman Rank, *rho*) between human judgments and similarity measures for 5 classes

Lang	Correlation with human judgments		Correlation between similarity measures
	Anchor+word overlap	Translation	
DE-EN	0.631	**0.703**	0.897
EL-EN	**0.124**	0.077	0.441
ET-EN	-0.045	**-0.001**	0.741
HR-EN	**0.495**	0.408	0.683
LT-EN	0.376	**0.512**	0.791
LV-EN	0.362	**0.497**	0.593
RO-EN	**0.279**	0.250	0.680
SL-EN	**0.417**	0.385	0.576

Table 7: Correlation (Spearman Rank, *rho*) between human judgments and similarity measures and between similarity measures for 5 classes and across languages

Table 7 shows the correlation of similarity scores with human judgments for each language pair. We observe that the correlation varies widely based on the language pair. For example, human judgments for the DE-EN language pair correlates highly with both measures of similarity; however the correlation for Estonian (ET-EN) is very poor. The anchor+word overlap

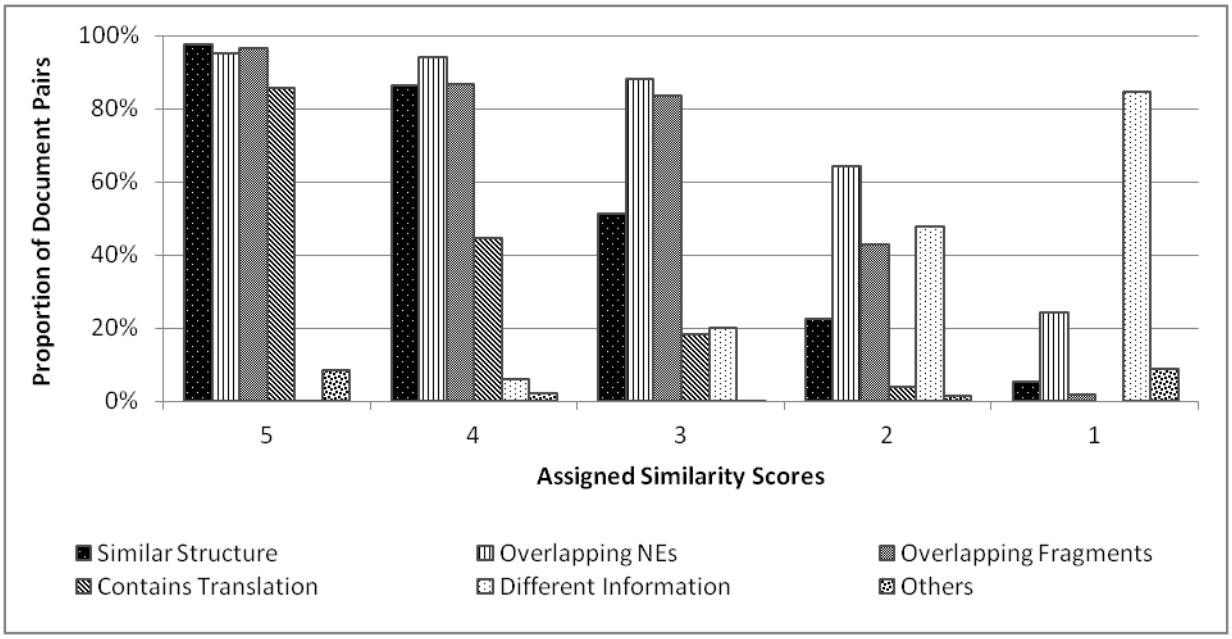

Figure 5: Characteristics that capture various levels of similarity.

measure of similarity has higher correlation than the translation approach with the human similarity judgments for 4/8 of the language pairs. This is a positive result given that the result is obtained using a language-independent approach making use of only a bilingual lexicon derived from Wikipedia. From Table 7 we also find that the correlation between the similarity scores overall is good but, again, varies depending on each language pair. For example, correlation is lowest for EL-EN which may suggest poorer MT results for Greek to English.

4.4 Classification Task

In this section we compare the two approaches for measuring similarity based on using the scores from each approach as features in a classification task. For each document pair, we round up the average assigned similarity scores to the nearest class, e.g. document pair with average score of 4.5 is included in class 5. Using a Naïve Bayes classifier[9] and 3-fold cross-validation, we are able to classify 40% of the 800 cases correctly using the scores from anchor+word overlap method (similar performance was achieved using the translation method). Taking the Most Common Class (class 5, *N=297*) as a baseline then simply assigning all cases to this would result in an accuracy of 37.1%. We find that many (36%) of the mis-classified cases are between classes 4 and 5. The classifier correctly classified 52.5%, 37.2% and 38.2% of document pairs in classes 5, 4 and 3, respectively. None of the document pairs in classes 1 and 2 were correctly classified; most probably due to the small number of available training documents (5 and 61 cases respectively). These document pairs were incorrectly classified as class 3 instead. Combining human judgments into 2 classes, as described in Section 4.2, we can correctly classify 58% of cases using either similarity score (this represents 50.2% of similar documents and 66.8% of non-similar documents). Accuracy for the Most Common Class baseline is 52.5% (for the class 'similar').

[9] In these experiments we used the Weka Toolkit (version 3.4.13).

5. Discussion

5.1 Features of 'Similar' Articles

As stated in the introduction, one of the goals of this study was to better understand what makes two Wikipedia articles written in different languages similar. The evaluation scheme has enabled us to analyse the characteristics of document pairs from each similarity score in more detail (Figure 5).

When judging cross-language similarity the judges were asked to provide input on what led them to make their decision. The options included whether the two articles contained a similar structure or ordering of the content (*similar structure*), whether documents contained overlapping named entities (*overlapping NEs*), whether fragments of text (e.g. sentences) from one document could be aligned to the other (*overlapping fragments*), whether content in one article appeared with equivalent translations in the other (*contains translation*), whether articles contained different information or from a different perspective (*different information*) and any other reasons. The results suggest that majority of document pairs judged as highly similar (either a score of 4 or 5) in Wikipedia have the following characteristics: they contain similar structure, overlapping named entities, overlapping fragments and over 80% of these document pairs contain what appear to be translations of the content, i.e. translation equivalents.

Interestingly the results also show that simply sharing named entities or having aligned segments of text

does not guarantee that the overall document pair is similar. The latter could be the result of judges making document-level similarity assessments: a document pair may contain a number of aligned sentences but at a document (or global) level the degree of similarity is low. As expected, the number of articles containing different information increases for little or non–similar cases (1-2). A distinguishing feature between text pairs which exhibit high similarity vs. those exhibiting little or no similarity would seem to be whether the content in the articles follows a similar structure and whether the document pairs contain translation equivalents of each other. To verify this, we created a binary feature vector for each of the human similarity judgments (1,600) and the comments (5 features in total) and performed feature selection using 3-fold cross-validation on the 5 classes of similarity judgment and a measure of information gain[10]. The features are ranked for their discriminative power in the following order: contains translation, similar structure, different information, overlapping fragments and overlapping NEs.

Using binary feature vectors based on the comments to classify all 1,600 cases we achieve an accuracy of 57% (the Most Common Class baseline accuracy is 31.4%). When considering a 2-class problem (high/low similarity) we obtain an accuracy of 81%. In this case, the Most Common Class accuracy baseline is 60.3%. This suggests that capturing these features of similarity could improve our measure of cross-language similarity.

Judges were also able to provide 'other' comments and several highlighted a number of non-English articles containing duplicate English sentences. It would appear that in these cases the assessors ignored the content during comparison; however, the computed measure of similarity would incorrectly count these cases as similar and thereby inflate the similarity scores. A solution to this would be to include a maximum threshold above which sentences are filtered out (similar to using a minimum threshold) or using language detection to detect such cases and ignoring them during the sentence alignment stage.

5.2 Measuring Cross-Language Similarity

A further goal of this study was to compare an adapted version of an existing method for cross-language similarity (Adafre & de Rijke, 2006) with an approach based on using freely available MT systems. In contrast with existing work we compare the computed similarity scores with human judgments to identify their degree of correlation. We also compare results obtained across a range of language pairs to determine the success of exploiting inter-language links in Wikipedia to develop a bilingual lexicon. A similarity score based on this approach seems to capture some essence of cross-language document similarity as judged manually. There are obvious weaknesses to our approach for some language pairs (e.g. ET-EN) that require further development. However, the issue is not resolved by using an MT system, which may simply reflect the difficulty faced when dealing with under-resourced languages that result in lower translation quality.

Through manual inspection we identified two cases where assessors disagree with the assigned anchor+word overlap scores: (1) assessors assigning a low similarity score to pairs which have high anchor+word overlap score; (2) assessors assigning a high similarity score to pairs which have low anchor+word overlap score. We found that the most common reason for the first case is that the shorter document (normally the non-English one) is a subset of the longer document. In these cases, documents are scored highly using the anchor+word overlap approach as the length of the smaller document is used to normalise the similarity score. Assessors, on the other hand, identified different information in the longer document not included in the shorter document. They therefore gave a lower similarity score.

In the second case, when similar documents are scored poorly using the anchor+word overlap approach, we find that one reason is due to the existence (or lack of) overlapping cognates. This results in better performance on languages with a similar written form to English, such as German. For other languages, such as Greek, the alphabets are very different and subsequently the number of matching cognates drops significantly. This then causes the approach to rely on the availability of links, which in some cases is not enough. There are similar documents which simply do not contain enough links for the language-independent method to identify parallel sentences accurately.

The findings also suggest that a lack of correlation in results is because the similarity of document pairs is assessed in different ways. The anchor+word overlap approach was initially developed to identify similar documents in Wikipedia for the purpose of building comparable corpora for MT. Therefore, the method is intended to identify Wikipedia documents which contain similar fragments. In situations described previously for case 1, documents are considered useful because they contain overlapping fragments, and therefore are scored higher. Whether or not the longer document contains a large amount of new information is irrelevant for the anchor+word overlap method. However, this does not relate very well to assessors' judgment, as they base their scores on the overall content of the Wikipedia articles. Therefore, further work is needed to better capture human similarity judgments.

6. Conclusions

In this paper we have analysed the performance of two similarity measures in identifying cross-language similarity between Wikipedia articles on the same topic but written in different languages. In this initial study, we evaluated 800 document pairs and found that similarity measures using machine translation and language-independent features based on mining anchor texts from

[10] We used weka.attributeSelection.InfoGainAttributeEval for feature selection and weka.attributeSelection.Ranker to rank the features from the Weka Toolkit.

inter-language links in Wikipedia correlate with each other (*rho=0.744, p<0.01*) and to a lesser degree with human judgments (*rho=0.353, p<0.01 and rho=0.325, p<0.01*). We have shown that our measure of similarity varies widely across language pairs with, for example, German-English results correlating better with human judgments than Estonian-English.

The performance of the language-independent method is comparable to an approach based on translating articles into English and determining similarity monolingually in several language pairs. This demonstrates the potential benefit of mining inter-language links from Wikipedia for under-resourced languages. We also show that the similarity measures can be used to perform classification with an accuracy of 40% for 5 levels or classes of similarity. We have also analysed features which judges have identified to capture cross-language similarity between articles and have used this analysis to uncover distinguishing features of the various levels of similarity.

There are several avenues to explore in future work. These include improving the cross-language similarity measures by incorporating term weighting rather than using binary feature vectors; automatically capturing the features identified by the judges to distinguish similar and non-similar document pairs and improving the sentence alignment algorithm to include cases when sentences are split up between the source and target documents.

7. Acknowledgements

The project has received funding from the ACCURAT Project, European Community Seventh Framework Programme (FP7/2007-2013) under Grant Agreement Number 248347. We also thank all of the 16 assessors from ACCURAT who judged document pairs and provided the human judgments.

8. References

Adafre, S. F. and de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In Proceedings of the EACL Workshop on New Text, Trento, Italy.

Bharadwaj, R. G and Varma, V. (2011). Language Independent Identification of Parallel Sentences using Wikipedia. In Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pp. 11–12.

Erdmann, M.; Nakayama, K.; Hara, T. and Nishio, S. (2008). Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia. Journal of Information Processing 16, pp. 67-79.

Filatova, E. (2009). Directions for exploiting asymmetries in multilingual Wikipedia. In Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3 '09).

Hatzivassiloglou, V.; Klavans, J. L. and Eskin, E. (1999). Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning, In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 203-212

Lu, Y.; Huang, J. and Liu, Q. (2007). Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In Proceedings of the 2007 EMNLP-CoNLL, pp. 343-350.

Munteanu, D. S. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Comput. Linguist: 31(4), pp. 477-504.

Lin, W.; Snover, M. and Ji, H. (2011). Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes. Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pp. 43–52, Edinburgh, Scotland, UK, pp. 27–31.

Otero, P. G. and López, I. G. 2010. Wikipedia as multilingual source of comparable corpora. In Proceedings of the LREC Workshop on BUCC, pp. 30–37.

Patry, A. and Langlais, P. (2011). Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pp. 87–95.

Smith, J. R.; Quirk, C. and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. InNAACL-HLT 2010.

Tomás, J.; Bataller, J. and Casacuberta, F. (2001). Mining Wikipedia as a Parallel and Comparable Corpus. In Language Forum, volume 1, pp. 34.

Yu, K. and Tsujii, J. (2009). Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. Proceedings of the NAACL-HLT 2009.

Zesch, T.; Műller, C.; and Gurevych, Iryna. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wikictionary. Proceedings of the LREC 2008, Marrakech, Morocco.