Supporting Collaboration in Wikipedia Between Language Communities

Ranjitha Gurunath Kulkarni
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, Pennsylvania 15213
ranjithagk@gmail.com

Gaurav Trivedi
National Institute of Technology Karnataka, Surathkal
Srinivasnagar, DK, Karnataka, India 575025
gtrivedi@nitk.ac.in

Tushar Suresh
National Institute of Technology Karnataka, Surathkal
Srinivasnagar, DK, Karnataka, India 575025
iamtushar@gmail.com

Miaomiao Wen
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, Pennsylvania 15213
mwen@cs.cmu.edu

Zeyu Zheng
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, Pennsylvania 15213
zeyuz@cs.cmu.edu

Carolyn Rose
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, Pennsylvania 15213
cprose@cs.cmu.edu

ABSTRACT
This paper describes an application of machine translation technology for supporting collaboration in Wikipedia. Wikipedia hosts separate language Wikipedias for hundreds of different languages. While some content is specific to these different versions of Wikipedia, some topics have pages within multiple different Wikipedias. Similarly, while some users participate only in one Wikipedia, we find users who play a bridging role between these sub-communities and participate in the process of maintaining similar pages in different Wikipedias. Since these are not the majority of users, a support tool that allows stretching the effort of these specialized users further by indicating where their effort is needed could be a tremendous benefit to the community. An evaluation of the proposed approach demonstrates promise that such a tool could substantially reduce the effort involved in playing this bridging role on Wikipedia.

Author Keywords
Wikipedia; Computer Supported Cooperative work; Cross-lingual Document Similarity.

ACM Classification Keywords
H.5.2; I.2.7

General Terms
Human Factors

ICIC'12, March 21–23, 2012, Bengaluru, India.

INTRODUCTION

In this paper we present a tool for supporting inter-cultural collaboration in Wikipedia. Wikipedia is valued worldwide as a repository for knowledge, with pages in over 280 languages. In 39 of these languages, it has over 100,000 pages, and at least 10,000 in 65 more languages. Not only does it provide a valuable resource for the individual cultures that possess their own language specific Wikipedia, it also provides a context in which a meaningful inter-cultural exchange can take place online. In particular, for many topics, similar pages exist in the Wikipedia of multiple languages, and some editors who possess the requisite language skills contribute towards maintaining different versions of the same page in two or more language specific Wikipedias.

For example, one of the most popular pages for September 27, 2011 on the English Wikipedia was a page about Wangari Maathai, who was Kenya's first woman to have earned a Nobel peace prize. Not surprisingly, there is a corresponding page about her in the kiSwahili Wikipedia. Some of the editors are shared between the two versions of the page. For example, on September 27, 2011, a user with username Lucas-bot made an edit to the English version of the page. And on the same day, the same user made two edits to the kiSwahili version of the page. Despite valuable bridge people between language specific Wikipedia communities, the multiple versions of the same page within different language Wikipedias are not kept equally up to date. For example, the kiSwahili version of the Wangari Maathai page is very short and mainly discusses the Nobel Peace Prize that she won, while the English version of the page is much more elaborate, giving much more detail about her life, travels, and educational experiences abroad in the US and in Germany, personal information about her private life, and in depth discussion of her political views.

Keeping content up-to-date in all of the language specific versions of the same page is challenging. Not all contributors to Wikipedia are capable of maintaining similar pages in multiple languages. Thus, in order to use human resources most efficiently, a solution that identifies those places where the contributors who possess this specialize knowledge are needed to perform this valuable community service would facilitate this inter-cultural exchange and thus benefit the international Wikipedia community as a whole. In this paper we present a technical solution for identifying these opportunities by employing machine translation technology to translate content from source pages into English, and then compare content in English to identify the extent to which inconsistencies exist. Due to the limitations of web accessible machine translation techniques, it is challenging to find similarities in text written in different languages. Along with directly comparing the text on English and corresponding translated non-English page, our approach takes into consideration, other language-properties like homophones and synonyms.

We focus on content included in the Wikipedia Info Boxes. Much prior work in the Machine Translation community has already targeted the problem of aligning texts from multiple languages [22, 23, 24] Semi-structured knowledge, such as is found in Info Boxes, presents different challenges. The terse nature makes it less amenable to existing text similarity measures [25, 26, 27], and the decontextualized nature of the encoded knowledge makes it less amenable to typical alignment techniques. Thus, our evaluation provides knew knowledge that contributes to a longer term effort to support inter-cultural collaboration in the Wikipedia community.

In the remainder of the paper, we first review some related work. Then we offer and overview of the proposed multi-stage technical approach. Next, as a proof of concept we present an ablation study in which we evaluate the contribution of each separate level towards the accuracy we achieve on a parallel English-Hindi corpus, with 50 pages represented both in the English and the Hindi Wikipedias. Next, we present a larger evaluation on all pairs of English, Chinese, and German on a parallel corpus of 100 pages represented in the English, Chinese, and German Wikipedias. We conclude with a discussion of our continued research.

PREVIOUS WORK

Several studies on Wikipedia highlighting the cross-cultural differences have been done in the past. Findings from such work show that there is a significant difference in various language wikipedias [17]. They support the view that Wikipedia is a diverse community where cultural differences that exist in the world external to the online Wikipedia world also exist within that online world.

Within this context, Wikipedia bots have proven to be a very useful resource to increase contributions from members [19]. This line of work builds on prior social psychology work on feeling of responsibility that demonstrates that people feel less responsible when they know responsibility is shared rather than specific to them [18]. Users who are able to contribute to multiple language specific Wikipedias are a valuable resource. While their abilities are not unique, they are unusual. Thus, the problem we target in this paper would fall within the scope of applicability of these findings. Thus, for our application, if it were possible to create a bot that could identify opportunities where these bridge people were needed to address mismatches in content, it could apply findings from this work by emphasizing the specialized knowledge that user had with respect to this identified maintenance need.

Studies on the pattern of growth in contribution to communities like Wikipedia [20] illustrate a trajectory in which novice users begin their editing experience by correcting minor mistakes in articles. As they move from peripheral participation into more core participation, they shift their concern from these relatively insignificant details to the maintenance of the integrity and overall quality of Wikipedia. Shifts in the nature of edits to Wikipedia pages over time may indicate movement along this trajectory and may provide useful information in identifying users to target with update requests. Furthermore, core Wikipedia community members may value tools that assist them with their goal of maintaining the integrity of Wikipedia by directing their effort where it is most needed. Such tools may be viewed as belonging to the set of affordances of the kind of socio-technological infrastructure that helps in maintaining the quality of Wikipedia [21].

Our work focuses on the problem of identifying the mismatches in content. This task requires comparing content across similar pages in different language specific Wikipedias. Existing techniques for document similarity [25, 26, 27] leverage patterns of word co-occurrence and word order regularities. In case of cross-lingual document similarity, assumptions such as these must be relaxed since, among other concerns, we know languages differ systematically in terms of word order. Automatic translation services that are freely or cheaply available today are weak in addressing such issues. Since our work leverages these weak translators, we must consider that the results we get as part of our process may be noisy and unreliable. Furthermore, we aim to evaluate how far we can push a simple and fast approach that can readily be applied to relatively low resource languages that nevertheless have web-based translation services available.

Prior work on cross-lingual document similarity compensates for some of these issues using carefully constructed multi-lingual thesauri, which are unfortunately only available in certain languages [2]. Relying on a resource like this would limit the generality of our approach to low resource languages such as Indian and African languages. In contrast, our proposed technique places only

limited constraints of the choice of language. Specifically, any language for which there is an automatic translation service available on the web is fair game. Google translate, as an example, provides a translation service for 63 languages to English at the time of writing this paper.

While our task is unusual, it is not completely unique. Information arbitrage across multilingual Wikipedia [16] is another related effort that has so far been demonstrated with four European languages, namely English, Spanish, French, and German. However, this prior work does not demonstrate any generality to languages outside of Germanic and Romance languages, which share a relatively similar language structure; whereas we have demonstrated generality of approach to Hindi and Chinese, which have a far more different structure from English.

TECHNICAL APPROACH

We describe an approach to checking the consistency of information contained in info boxes on pairs of pages devoted to the same topic but in different language Wikipedias. We first give an overview of the process and then describe the technical details of the most complex portions of the process in greater depth.

Overview

The example in Figure 1 illustrates an example where the information contained on corresponding pages in two Wikipedias is very inconsistent, and both could benefit from including information contained on the alternate page. Specifically, we see the info box from the Barclays Bank page on the English and French Wikipedias respectively. The English version focuses on the history, assets, services, and trade related information. The French page, in contrast, contains information necessary to make international bank transfers, such as the BIC and IBAN.

Figure 1. Example of Info boxes extracted from the Barclays Bank page on the English Wikipedia (left) and the French Wikipedia (right).

An overview of the proposed translation and matching process can be found in Figure 2. One goal of the approach is to apply to as wide range of language pairs as possible. Thus, we choose English as an intermediate representation. Regardless of the source language of a page, we first extract attribute value pairs from all of its info boxes. For pages in source languages other than English, these attribute value pairs are then translated into English using an online automatic translation service, specifically the Google Translate API. These translation services are known to produce errorful translations. However, part of the contribution of the work presented in this paper is a demonstration that nevertheless, we can use the output from these online translation engines usefully in our process.

The English attribute value pairs from both pages are then passed into an Attribute Name Pairwise Matching module that identifies potential matches between attribute names on the two pages. This is nontrivial since in addition to potential errors introduced at the translation stage, the translation may result in attribute names from the two pages having the same meaning but expressed through different words. When a potential match is identified, then the attribute values must be matched to determine whether the information contained within the corresponding info box entries is consistent across the two pages. This matching process is challenging for similar reasons to those just enumerated for the attribute name matching stage.

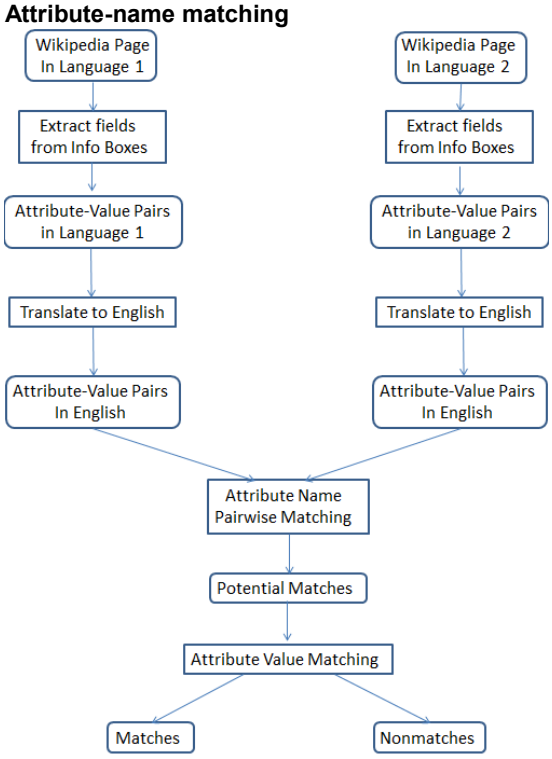

Figure 2. Overview of translation and matching process

The attribute-names comparison is done in two levels. An initial very conservative comparison is done by computing the percentage similarity between the two strings (or sets of

strings) using the similar-text function provided by the standard PHP library [9, 10]. The pairs of attribute-names (e.g., of English and Hindi) are marked as a match if their percentage similarity is greater than a threshold. In this case, we set the threshold to 95% so as to mark only those pairs which match to a very high extent. At this stage, most of the matching attribute names are identical or nearly identical.

The attribute names from either page that did not find a match in the first stage are then fed into a second, less conservative matching process. To overcome some of the challenges and limitations of the automatic machine translation, we introduce sense matching into our approach. For each word in the attribute name, sets of synonyms called synsets are obtained from a language resources called WordNet[12]. Using the synsets obtained from WordNet, we can compute the percentage of overlap between the synsets extracted for every pair of attribute-names. A normalized score is obtained for each attribute-name pair (A, B) using the following formula:

$$\text{Score (A, B)} = \frac{(Words\ with\ matching\ sense)^2}{(words\ in\ A) * (words\ in\ B)}$$

All those pairs with a score greater than a threshold are marked in the comparison matrix just as in the previous stage. In case of more than one match for a given attribute-name, the one with the highest match score is selected. Examples of attribute names that match include things like *Altitude* and *Height* or *Ruler_position* and *Leader_title*.

Attribute-value matching

Once matching attribute names are found across pages, the associated values must then be checked for consistency. To obtain a similarity assessment between attribute values, as mentioned above, the first step is a direct comparison where the set of letters and their ordering is considered. As most of the times the attribute values are either nouns or names, there is a high possibility that the translator may get the spelling wrong. Therefore the next step is to check if the two values are homophones. We first identified nouns using a Parts Of Speech tagger [15]. We then generated metaphone codes [13][14] for each of the identified nouns, and then checked for equality. If the metaphone codes match, then the pair of values was considered as homophones and a match was declared. An example of metaphones declared as a match is "*bayaluseeme*" and "*bayalusime*". Every comparison of attribute value pairs is associated with a normalized score. All attribute pairs evaluated with a score less than the threshold are reported to be inconsistent. Additionally, attributes that are missing on either of the articles are flagged as inconsistent. Examples of inconsistences are values of "5th" and "3rd" for population rank or "*siddaiah, ias*" and "*dr. r. subrahmany*" for leader name.

EVALUATION

We conducted two separate evaluations of our approach. The first evaluation was designed to evaluate the contribution of each stage in the process separately, in order to motivate the design of the approach. The second evaluation was conducted on a larger set of examples, but only examines the final result rather than evaluating the output at each stage in the pipeline. We consider the first evaluation to be a proof of concept, and the second to be a larger scale, fully fledged system evaluation. Both evaluations demonstrate that the pipeline is well motivated and that the approach, though simple, holds promise for assisting in keeping the variety of existing language Wikipedias consistent with one another.

Datasets

We constructed a dataset for each evaluation by finding corresponding pages represented within 2 or 3 language Wikipedias where the pages within the resulting set all contained info boxes.

Data set 1

The first set comprised a collection 50 English and corresponding Hindi Wikipedia articles on the cities of India. We observed that pages within this scope had a lot of inconsistencies between their corresponding articles in different languages. The 50 cities were selected randomly from the set of page pairs that met the criteria just mentioned.

Data set 2

The second set comprised 100 Wikipedia articles each in English, German and Chinese on US based companies. These articles were chosen based on the availability of infoboxes in all the three language Wikipedias.

In both evaluations the same pipeline illustrated in figure 2 was applied to identify inconsistencies between all pairs of languages. Thus, for the second evaluation, 3 pairwise comparisons were made, specifically English-German, English-Chinese, and Chinese-German.

Evaluation 1: Proof of Concept

As mentioned, the first evaluation was meant as a proof of concept that demonstrates the contribution of each stage within the pipeline introduced earlier in the paper. The infoboxes of the collected dataset were compared using that pipeline, and inconsistencies were identified and noted using this fully automatic approach. We then calculated the following performance metrics: Accuracy, Precision, Recall and F-measure. These are standard metrics that allow us to assess the quality of automatic matching our approach achieves. These metrics were computed by comparing the automatic analysis to a gold standard analysis created by native speakers of the languages involved. This gold standard was constructed as follows.

First the judges examined the full set of attribute value pairs from the two pages. Each one was then assigned to one of the following two categories:

E1: Attributes that are classified as present and similar

E2: Attributes that are either missing or dissimilar

For each attribute, we can also obtain a classification into E1 or E2 based on the system's assessment of match or non-match on attribute names, and then consistent versus inconsistent on the values of the matching attributes.

Gold Standard: Manual classification of attributes.

		Condition as determined by the *Gold standard*	
		true	**false**
Obtained classification	**E1**	**tp** (true positive)	**fp** (false positive)
	E2	**tn** (true negative)	**fn** (false negative)

Figure 3. Confusion matrix used to compute precision, recall and f-measure from the gold standard and automatically assigned E1 and E2 codes.

Precision, recall, and f-measure of result was computed for three versions of the system. First, a Baseline system simply evaluated exact match on attribute names and values. This is the most conservative and simple approach. Next, a Level 1 system included Wordnet based synset matching, but not homophone matching. A final Level 2 system included also the homophone matching.

	Accuracy	**Recall**	**F-measure**
Baseline	0.76	0.29	0.45
Level 1	0.79	0.38	0.54
Level 2	0.85	0.58	0.72

Table 1. Results of Evaluation part 1

Because even the Level 2 system was very conservative in identifying matches, there were never false positives in this evaluation, and thus the Precision is always 1. For the reported metrics in Table 1 we see consistent increases in performance as we move from the most conservative baseline system to the least conservative system.

The increase in recall at every level was due to the decrease in the number of false negatives. From baseline to level 1, the number of attribute names which were wrongly declared as unmatched were now being matched. Ex: Consider the attribute-name-value pair,

*Area = Bayalusime (*Hindi*)*

*Region = Bayaluseeme (*English*)*

Here, the term Area if translated back to the native language Hindi, actually refers to Region. But due to translation errors, this attribute-name pair comes out as a mismatch in the baseline system. Whereas, after the synset matching is introduced in level 1, such errors are minimized and hence the above argument of increase in recall.

Similarly in case of level 1 to level 2, some attribute names that were previously labeled as unmatched were found to match. This can be seen in the above example, where the noun "*Bayaluseeme*" has different spellings but sound the same. Such attribute values get matched at level 2 which checks for homophones. Hence the further increase in recall.

In order to interpret what the performance values mean in terms of human effort, we developed an additional performance metric, which we have termed "Reduced human effort". This metric computes the amount of work saved by humans by the system's work in flagging the inconsistencies. This metric is defined as follows:

$$\textit{Reduced Human effort} = \left\langle \frac{X}{Y} \right\rangle$$

Here, X is the number of inconsistencies found on a page, while Y denotes total attributes present on the page. The reduced human effort calculated showed us that on an average, around 55% of the set of attributes were either missing or inconsistent on either of the two infoboxes. By flagging these, a human user who is capable of fixing the info boxes on the page pairs could save time by focusing attention on the subset that was identified automatically rather than checking everything manually.

In future work we may experiment with lowering the threshold values to make the process less conservative. In that case, we expect that recall will increase and precision will decrease. Currently, however, we find it more advantageous to maintain the conservative stance since it eliminates the problem of false positives. We are more interested in those attributes that are wrongly classified as a mismatch and will be moved to their right classes after application of our approach.

Evaluation 2: A Larger Multi-Lingual Evaluation

As the second dataset comprised of 3 languages, evaluation was done on 3 pairs of languages namely, English- German, Chinese- German and English- Chinese.

They were evaluated at two levels:

Level 1: In this analysis, only the matching of Attribute-name pairs across languages is evaluated. Consider that there are X attributes on the first page and Y attributes on the second page. Each of the X attributes on the first page can be matched with any of the Y attributes on the second page. Furthermore, each attribute on one page has the possibility of not being matched to anything on the other page. Thus, the total number of pairwise decisions that need to be made about matches is (X+1)*(Y+1)-1. Out of those, some subset M should be marked as a match based on an expert judge's gold standard analysis. Some subset N was marked as a match by the system. Thus, Precision is computed as the cardinality of the intersection between M and N divided by the cardinality of N. And Recall is the cardinality of the intersection of M and N divided by the cardinality of M.

Level 2: In this analysis, the number of Attribute-value pairs that matched is evaluated (given the number of attribute-name matches in level 1). Thus, we began with the pairs that were identified as matches in the Level 1 analysis regardless of whether it was correct. There were several cases to consider:

- If an attribute from one page was correctly identified as not matching any attribute from the other page, it was counted as a true negative.
- Attribute pairs that should have been identified as matching but were not were counted as false negatives.
- If an attribute pair was incorrectly matched, it was counted as a false positive in this analysis.
- If an attribute pair was correctly matched and the values were correctly identified as a match, it was counted as a true positive.
- If an attribute pair was correctly matched and the values were correctly identified as not matching it counted as a true negative.
- If an attribute pair was correctly matched but the values were incorrectly identified as not matching it was counted as a false negative.
- If the attributes were correctly matched and the values were incorrectly identified as matching it was counted as a false positive.

Precision, Recall and F-measure were calculated and the following results were obtained:

	Precision	Recall	F-Measure
Level 1	0.63	0.55	0.57
Level 2	0.86	0.63	0.70

Table 2. Results for English-German pair

	Precision	Recall	F-Measure
Level 1	0.95	0.77	0.84
Level 2	0.74	0.62	0.64

Table 3. Results for English-Chinese pair

	Precision	Recall	F-Measure
Level 1	0.72	0.45	0.54
Level 2	0.74	0.69	0.69

Table 4. Results for German-Chinese pair

The results obtained in the second stage of evaluation record high values of Precision, Recall and F-measure (at level 1) for the Chinese - English language pair when compared to the other two pairs. This can be accounted by the fact that both Chinese and English Wikipedia infoboxes have very similar attribute names even before translation (template given in English) unlike German infoboxes (as illustrated in following figures 2, 3, and 4).

```
{{Infobox dot-com company
| company_name     = Amazon.com, Inc.
| company_logo     = [[File:Amazon.com-Logo.svg|250px]]
| company_type     = [[Public company|Public]]
| traded_as        = {{Nasdaq|AMZN}}<br>[[NASDAQ-100|NASDAQ-100 Component]]
| foundation       = 1994
| founder          = [[Jeff Bezos]]
| location         = [[Seattle]], [[Washington (state)|Washington]], U.S.
| area_served      = [[World]]wide
| key_people       = [[Jeff Bezos]]<br><small>([[Chairman]], [[President]]
```

Figure 3. Infobox Template of an English article (screenshot from Wikipedia edit page)

```
{{Infobox Company
| company_name     = 亚马逊公司
| company_logo    = [[File:Amazon.com logo.svg|225px]]
| company_type     = [[上市公司]] ({{nasdaq|AMZN}})
| company_slogan   = "...and you're done"
| foundation       = 1994
| founder          = [[:en:Jeff Bezos|Jeffrey P. Bezos]]
| location         = {{flagicon|USA}} [[西雅图]]
| area_served      = [[全球]]
| key_people       = [[:en:Jeff Bezos|Jeffrey P. Bezos]]<br /><small>([[主席]])
```

Figure 4. Infobox Template of a Chinese Article (screenshot from Wikipedia edit page)

```
{{DISPLAYTITLE:amazon.com}}
{{Dieser Artikel|behandelt das Unternehmen ''Amazon''. Zu anderen Verwendunger
{{Infobox Unternehmen
| Name             = Amazon.com, Inc.
| Logo             = [[Datei:Amazon.com-Logo.svg|180px|Logo]]
| Unternehmensform = [[Gesellschaftsrecht der Vereinigten Staaten#Corporation|
| ISIN             = US0231351067
| Gründungsdatum   = 1994
| Sitz             = [[Seattle]], [[Vereinigte Staaten]] {{USA|#}}
```

Figure 5. Infobox Template of a German Article (screenshot from Wikipedia edit page)

Hence when German infoboxes are compared, the attribute names are translated into English, which introduces a large number of translation errors. For example, "*Sitz*" in German refers to "*Location*" in English/Chinese articles. But it gets translated as "*Seat"* which even after comparing synsets will not match with "*Location*".

The values of evaluation metrics increase from level 1 to level 2 in case of German-English and German-Chinese pairs. This is because the comparisons to be made after the attribute names match mainly include numbers and nouns. Hence only the conservative text comparison and homophones matching are needed to give us these results. But there are some errors as explained in the next section, which if avoided might lead to better results.

The decrease in the values in case of the Chinese-English pair is because of translation errors getting introduced at this level. The attribute values in Chinese articles are all in Chinese though their attribute name were in English in the template.

Errors that contribute to the system can be divided into two main categories: 1. Syntactic errors. 2. Semantic errors.

Syntactic errors refer to the errors introduced in the system due to errors in the way the text is represented:

For example, attribute names that are represented as a single token in English may be represented as a phrase after translation from German, and that phrase may have a nontrivial match with the corresponding English tag. This occurred at least once in approximately 20% of the pages.

Semantic errors are more common compared to the syntactic errors. Here are a few of them,

1. Attribute-values in one language article being a single word description for their corresponding set of words in another language.

Ex:

*industry=graphics card, motherboard, power supply, desktop computer and pc accessory manufacturing (*German*)*
*industry=computer hardware (*English*)*

2. There might be composite attributes (one attribute in one language article may be a combination of two or more attributes in another language article.)

Ex:

"Foundation" in some articles in Chinese gives us information about both "*date*" when the company was founded and its "*location*". Whereas English has different attributes for "*date*" and "*location*".

3. Sense disambiguation errors because of bugs in translator. This is the most common type of error seen. Average number of such cases seen was 2-3 per article of 9-10 attributes in case of German to other language comparisons. This error is not evident in English – Chinese comparison as the attribute-names do not undergo translation.

English Spanish French

```
{{Infobox_Unternehmen
| Name           = Autodesk, Inc.
| Logo           = [[Datei:Autodesk logo.svg|200px|Logo von Autodesk]]
| Unternehmensform =
[[Gesellschaftsrecht_der_Vereinigten_Staaten#Corporation|Incorporated]]
| ISIN           = US0527691069
| Gründungsdatum  = 1982
| Sitz           = [[San Rafael (Kalifornien)|San Rafael]], [[Vereinigte Staaten|USA]]
| Leitung        = [[John Walker (Programmierer)|John Walker]], Gründer<br /> [[Carl
Bass]], [[Chief Executive Officer|CEO]] und Präsident
| Mitarbeiterzahl  = > 6.800
| Branche        = Softwarehersteller
| Umsatz         = $ 1.952 Milliarden [[US-Dollar|USD]] (FY2011) <ref>
[http://www.wikinvest.com/stock/Autodesk_(ADSK)/Data/Income_Statement#Income_St
atement Umsatz]</ref>
| Homepage       = [http://www.autodesk.com/ www.autodesk.com]
}}
```

English Spanish Arabic

```
{{Infobox_Unternehmen
| Name = Autodesk, Inc.
| Logo = [[File: Autodesk logo.svg | 200px | Logo Autodesk]]
| Corporate form = [[# Gesellschaftsrecht_der_Vereinigten_Staaten Corporation |
Incorporated]]
| ISIN US0527691069 =
| Founded = 1982
| Seat = [[San Rafael (California) | San Rafael]], [[United States | USA]]
| Line = [[John Walker (programmer) | John Walker]], founder <br /> [[Carl Bass]], [[Chief
executive officer | CEO]] and President
| = Number of employees> 6,800
| Industry = Software Vendor
| Revenue = $ 1.952 billion [[USD | USD]] (FY2011) <ref>
[http://www.wikinvest.com/stock/Autodesk_ (ADSK) / Data / Income_Statement #
Income_Statement Sales] </ ref>
| Homepage = [http://www.autodesk.com/ www.autodesk.com]
}}
```

Figure 6. Illustrating Errors due to Translation from German to English (screenshot from http://translate.google.com)

CONCLUSION

In this paper, we present an approach to automatically point out differences between two articles written in different languages. We proposed an approach that uses the concepts of homophones and synonyms in addition to direct comparison. Our evaluation showed that there was a significant increase in recall after the concepts of homophones and synonyms were applied in addition to the direct text comparison.

Two evaluations on two different language sets that cumulatively include English, German, Chinese, and Hindi, which have very different language structures, demonstrates that the proposed approach has some generality across languages. It can also be seen that the two domains considered namely Cities of India and Companies based in the United States were not similar and hence our method also has some generality across domains where the corresponding pages in Wikipedia include info boxes.

The high number of inconsistent and missing attributes suggests that there is need for such automation and a bot that could leverage such analysis might be a useful tool to support inter-cultural collaboration in the Wikipedia community.

Nevertheless, weaknesses in the results suggest that the simplistic approach taken here may be substantially improved using more sophisticated language analysis techniques. Our approach currently depends on the lexical database "WordNet" for synonyms. We anticipate that stronger lexical databases would improve the effectiveness, potentially substantially increasing the recall. The current method does not address some common types of translation errors namely phrase translations. When phrases that are translated from one language to another are compared with single words giving the same meaning as the phrase, our system fails to recognize the match. One reason is that WordNet does not provide any information about how words may be paraphrased into a short phrase. Also, abbreviations, units conversion and geographic location matching is not handled by the current system.

To address the above mentioned errors, there are a few solutions we are focusing on, namely: for geographic location matching, the use of Gazetteer databases might help in providing information about different names and formats in which a given region is referred to. The abbreviations issue can be handled by introducing domain specific dictionaries for abbreviations. The template issues mentioned earlier in the paper need special attention as well. Thus we plan on improving the system by addressing all of these issues one by one.

In the long term, our goal is to expand on this work and produce a Wikipedia bot that can be used to support the work of bridge editors between similar pages on separate language Wikipedias. The bot would highlight those parts of the page that need attention. In case, there is missing information, the bot would prompt them on that particular page. Also the bot would give information about which language version has the most updated information. Before this vision can be realized, however, the work presented in this paper must be integrated with approaches to text similarity [25, 26, 27] that would allow the technique to be generalized from info boxes to the main text of the article.

In reference to the scenario described at the beginning of the paper, the resulting bot would point out to a user seen as an editor both of English and KiSwahili pages that there is a lot of information missing in the Kiswahili version of Wangari Maathai page and that the user can refer to the English page to update it. Specific pointers both to the info boxes as well as the main text would be given. The targeted pointers would facilitate efficient intervention of the contacted user.

ACKNOWLEDGMENTS

This work was funded in part by National Science Foundation Grant SBE 0836012.

REFERENCES

1. Christof M ller and Iryna Gurevych. 2009. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval Evaluating Systems for Multilingual and Multimodal Information Access, Springer Berlin /Heidelberg, pp. 219-226.
2. Steinberger, Ralf and Pouliquen, Bruno and Hagman, Johan 2002. Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC EProceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pp. 415-424.
3. Aminul Islam and Diana Inkpen. 2008, Jul. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity ACM Transaction on Knowledge Discovery from Data, Vol. 2, No. 2, Article 10.
4. Wikipedia Infoboxes Help. (2010, Dec.) [Online].Available: http://en.wikipedia.org/wiki/Help:Infobox
5. Wikipedia Infoboxes Categories. (2010, Dec.)[Online].Available http://en.wikipedia.org/wiki/Category:Infobox templates
6. MediaWiki API Documentation. (2010, Dec.) [Online]. Available: http://www.mediawiki.org/wiki/APIox
7. GoogleTranslate API, developer's guide (v2): Using REST. (2010, Dec.) [Online]. Available: http://code.google.com/apis/language/translate/v2/using rest.html
8. Libcurl - C API documentation. (2010, Dec.) [Online]. Available: http://curl.haxx.se/libcurl/c/

9. PHP similar text function documentation (2010, Dec.) [Online]. Available: http://php.net/manual/en/function.similar-text.php
10. Jonathan J. Oliver. 2008, Jul. Decision Graphs - An Extension of Decision Trees. Available: http://www.cs.monash.edu.au/ jono/TechReports/TR173.dgraph.ps
11. Metzler, Donald and Dumais, Susan and Meek, Christopher 2007. Similarity Measures for Short Segments of Text Advances in Information Retrieval Vol. 4425,Springer Berlin / Heidelberg, pp. 16-27.
12. C. Fellbaum. 1998. WordNet: An Electronical Lexical Database. The MIT Press, Cambridge, MA.
13. PHP metaphone code generation function by Lawrence Philips. (2010, Dec.) [Online]. Available:http://php.net/manual/en/function.metaphone.php
14. Binstock & Rex. 1995. Practical Algorithms for Programmers Addison Wesley.
15. Parts Of Speech Tagging, PHP/ir, Information Retrieval and other interesting topics. (2010, Dec.) [Online].Available: http://phpir.com/part-of-speech-tagging
16. Adar, Skinner and Weld 2009, Information Arbitrage Across Multi-lingual Wikipedia WSDM'09, Barcelona, Spain
17. Ulrike Pfeil, Panayiotis Zaphiris, Chee Siang Ang 2006, Cultural Differences in Collaborative Authoring of Wikipedia.
18. B. Latane, K. Williams, and S. Harkins. Many hands make light the work: The causes and consequences of social loafing. J. Pers. Soc. Psych., 37:822–832, 1979.
19. D Cosley, D Frankowski, L Terveen… - 2007, SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia.
20. SL Bryant, A Forte… - 2005, Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia
21. Slattery, S. P. (2009). "Edit this page": the socio-technological infrastructure of a Wikipedia article. In *Proc. of the 27th ACM international conference on Design of communication* (pp. 289-296). Bloomington, Indiana, USA: ACM.
22. Liu, Y., Liu, Q., & Lin, S. (2006). Tree-to-string alignment template for statistical machine translation, *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*
23. Gildea, D. (2003). Loosely tree-based alignment for machine translation, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*
24. Och, F. & Ney, H. (2000). Improved statistical alignment models, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics.*
25. Mohler. M. & and Mihalcea, R. (2009). Text-to-text Semantic Similarity for Automatic Short Answer Grading, in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2009),* Athens, Greece
26. Gbrilovich, E. & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing, *Journal of Artificial Intelligence Research* 34(1).
27. Metzler, D., Dumais, S., & Meek, C. (2007). Similarity Measures for Short Segments of Text, *Advances in Information Retrieval,* Volume 4425, pp 16-27.