Providing Cross-Lingual Editing Assistance to Wikipedia Editors

Ching-man Au Yeung, Kevin Duh, and Masaaki Nagata

NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun
Kyoto, 619-0237, Japan
{auyeung,kevinduh}@cslab.kecl.ntt.co.jp, nagata.masaaki@lab.ntt.co.jp

Abstract. We propose a framework to assist Wikipedia editors to transfer information among different languages. Firstly, with the help of some machine translation tools, we analyse the texts in two different language editions of an article and identify information that is only available in one edition. Next, we propose an algorithm to look for the most probable position in the other edition where the new information can be inserted. We show that our method can accurately suggest positions for new information. Our proposal is beneficial to both readers and editors of Wikipedia, and can be easily generalised and applied to other multi-lingual corpora.

1 Introduction

There are currently over 250 different language editions in Wikipedia. However, significant differences exist between different editions in terms of size and quality [6]. Several projects on Wikipedia have been initiated to bridge this information gap with the help of both human and machine translation [12,13,14]. Google also provides a translator toolkit that assists users to translate Wikipedia articles[1].

While existing efforts focused on translating whole articles, we believe maintaining existing articles across different languages is also a major challenge. Wikipedia is by no means a static encyclopedia. Articles are constantly being revised by editors. As different language editions are being developed separately, it is likely that different language editions will contain different information, depending on the focuses of the editors or interests of the respective community.

Although Wikipedia is not intended to be an encyclopedia in which different language editions are exact translations of one another [14], it is desirable to keep any article up-to-date and comprehensive. However, the effort required to identify what should be translated can be prohibitively expensive, especially when the target document already has substantial content. This requires editors to continuously monitor articles in different languages, which is clearly unscalable.

We propose a framework that assists Wikipedia editors or translators to transfer information from one language into another. We term this task **cross-lingual document enrichment**. Our proposed framework is completely automatic and

[1] Google Translator Toolkit: http://translate.google.com/toolkit

A. Gelbukh (Ed.): CICLing 2011, Part II, LNCS 6609, pp. 377–389, 2011.

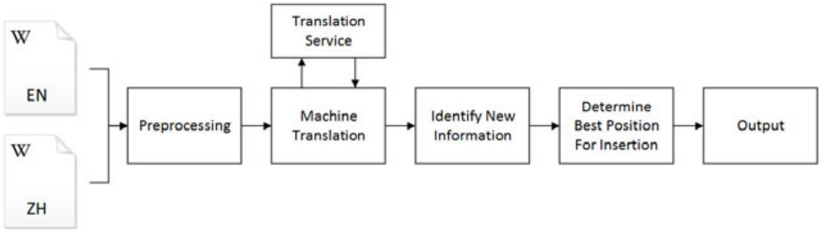


Fig. 1. System design of our proposed cross-lingual document enrichment framework

only requires the availability of a machine translation service. While we focus on Wikipedia in this paper, our techniques can be applied to any multi-lingual corpus where disparity of information in different languages is a problem. We believe this research has a positive impact on the creation and maintenance of huge multi-lingual corpora which have become more common nowadays.

2 Cross-Lingual Document Enrichment

While our proposal is independent of the languages involved, for concreteness of presentation, we assume that our source document is in English and the target document is in Chinese. We choose to treat *sentences* as the basic units that carry information. The two major processes in our proposed framework are:

1. **New information identification**: Given two sets of sentences (in Chinese and English), identify a subset (of English sentences) that contains information not found in Chinese. (Section 2.2)
2. **Cross-lingual sentence insertion**: Determine the best position where a translation of the new sentence should be inserted into the Chinese document, respecting the document's existing discourse structure. (Section 2.3)

Figure 1 depicts the overall system design of our framework. For each article, English and Chinese editions are preprocessed to remove formatting information. Sentences are extracted and labelled by section and paragraph IDs. To compare sentences in different languages, we make use of a machine translation tool[2]. We translate all Chinese sentences into English, so that the information content could be compared. However, in practice any process that maps the two editions to the same symbol set is possible. For example, we can translate the English to Chinese, translate both editions to French/Italian/Spanish, or any combination of the above methods[3].

[2] We use Google Translate `http://translate.google.com/` in this work but in theory any broad-coverage translation service is possible.

[3] In fact, we can also translate the two editions to a latent mapping that is not reminiscent of any human language, using machine learning techniques like [3].

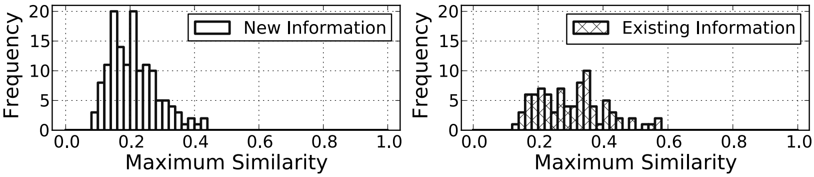


Fig. 2. Distribution of maximum similarity values of sentences with new or existing information for the article 'Angkor Wat'

In this paper, we use 'article' to refer to a particular topic in Wikipedia, such as 'Angkor Wat' or 'India'. An article has one or more language editions. We refer to each edition as a document. Let E be the document in English and C be the document in Chinese. We define a document as a sequence of sentences. Hence $E = (e_1, e_2, ..., e_M)$ and $C = (c_1, c_2, ..., c_N)$, where M and N are the numbers of sentences in English and Chinese respectively.

2.1 Measuring Sentence Similarity

To measure sentence similarity, we first submit the Chinese edition to a machine translation service and obtain an English translation. Then, for any document D, we extract a vocabulary V_D after stop-word removal and stemming. Each sentence s is represented by a term vector $s = (w_1, w_2, ..., w_{|V_D|})$, where w_i is the weight of the word $v_i \in V_D$ in s, determined by TF-IDF. The similarity of two sentences is calculated by the cosine similarity $\cos(s_i, s_j)$ between the two vectors. In our implementation, we add terms appearing in section titles to the term vectors of the sentences in the corresponding sections. We find that section titles are indicative of the topics of the sentences, and are helpful in improving the similarity metric.

2.2 Identifying New Information

To determine whether an English sentence contains new information with respect to the Chinese edition, we consider the following two methods.

Heuristic Method. Intuitively, English sentences with existing information should have high similarity to at least one sentence in the Chinese edition, while those with new information should have low similarity to all Chinese sentences. A heuristic method is to rank the English sentences by their maximum similarity to any Chinese sentence ($\max_j \cos(e_i, c_j)$), and consider sentences with maximum similarity lower than a certain threshold as containing new information.

Figure 2 shows the maximum similarity values for the article 'Angkor Wat'. The extreme values are relatively well separated across positive and negative samples, and the heuristic method will provide correct answer to a certain extent. However, there are regions where both positive and negative samples can be found. We suspect that this is due to limitations in the machine translation process. Thus, we also consider the following more sophisticated method.

Classification by Machine Learning. Alternatively, we can define this task as follows: given an article with English sentences $(e_1, e_2, ..., e_M)$, label each sentence e_i with $\{+1, -1\}$ where $+1$ indicates that the sentence contains new information and -1 otherwise. To avoid requiring any manual labelling effort, we adopt a *self-training* (see [1] and references therein) approach to classification.

We first order the sentences $(e_1, e_2, ..., e_M)$ by their maximum similarity, and choose the top $N\%$ of sentences as seeds for negative labels, and the bottom $N\%$ as seeds for positive labels. These labels are used to train a support vector machine (SVM) classifier. In this way, no manual annotations are required. The key assumption is that the extreme values of the maximum similarity are relatively reliable indicators of the true label. Based on our observations in Figure 2, we believe that the self-training assumption is reasonable on this kind of dataset.

We introduce several varieties of features for the SVM classifier. A feature vector is defined for each sentence e_i. The main types of features are:

- **Similarity**: Maximum cosine similarity of e_i. This is the feature used in the heuristic baseline (Section 2.2).
- **Neighbour**: Maximum cosine similarity of the neighbours, e_{i+1} and e_{i-1}. The idea is that if the neighbours have low similarity, then more likely e_i will contain new information, and the opposite is also likely to be true.
- **Entropy**: Entropy of similarity values of e_i, where similarity distribution is converted into probability distribution by $p(e_i|c_j) = \frac{\cos(e_i,c_j)}{\sum_{j'} \cos(e_i,c_{j'})}$. This feature counteracts situations where particular words lead to high cosine values for all sentences. Intuitively, an English sentence (if it contains existing information) should only be matched to a small number of Chinese sentences, and would achieve low entropy.

In practice, we have a total of 18 features, where each feature is a variant of one of the three main types listed above. For example, one feature is the entropy, while another (related) feature is the difference with the entropy averaged over the document.

2.3 Cross-Lingual Sentence Insertion

Our next task is to identify the positions in the Chinese edition where the sentences should be inserted. In some cases there may not be a single correct position for a new sentence as it can simply be inserted into a particular paragraph or section where the content matches that of the sentence. However, in other cases a sentence may elaborate an existing sentence and should be placed after that sentence. To accommodate this stricter requirement, we formulate our problem as finding a sentence c_j in the Chinese edition after which (translation of) the new sentence e_i should be inserted. To solve this problem, we consider several different methods as described below.

Insertion by Manual Alignment. As a first step, we consider that some sentences in English have been aligned to those in Chinese manually. Intuitively, the sentence should be inserted in a way that maintains the order of description

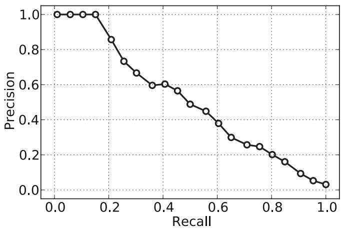


Fig. 3. Precision-recall curve of the similarity-based alignment for the article 'Angkor Wat'. Precision is recorded for every 0.05 recall level.

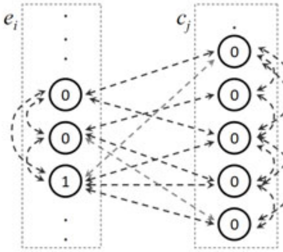


Fig. 4. A graph constructed based on the document structure and similarity values between sentences in two documents

or the flow of the article. Thus, a reasonable scheme is as follows. We look for an English sentence before e_i, say e_{i-1} that is manually aligned to a Chinese sentence c_j. Since e_{i-1} corresponds to c_j, it becomes natural that e_i when translated into Chinese should follow c_j as well. If e_{i-1} has no corresponding sentence in the Chinese edition, we can repeat the process and check e_{i-2} and so on. In practice, however, there will probably be no manually aligned sentences available for us to carry out this scheme. Hence, this method will mainly be used for comparison in evaluating our other proposed automatic methods.

Insertion by Similarity-based Alignment. When sentences are not manually aligned, we can generate some alignments by selecting pairs of sentences that achieve high values of similarity. Depending on the quality of translation, these pairs are very likely to be correct alignments. For example, after ranking the sentence pairs by their similarity values, we can heuristically choose the top 100 pairs as correct alignments. Figure 3 shows a typical precision-recall curve of one of the articles we manually aligned. In this figure, choosing the top 100 pairs corresponds to 47% precision and 55% recall, which is a well-balanced operating point. With these alignments, we can then apply the same method described in the above section. The limitation of this relatively simple method, of course, is that the sentence alignments may not be correct, leading to erroneous sentence insertions. In addition, highly similar sentences might be concentrated in a particular part of the article (e.g. the introductory sections).

Label Propagation. In view of the limitations of the above methods, we propose a method that is based on the technique of label propagation in classification and takes advantage of all similarity values among the sentences. Label propagation [17] uses a graph to incorporate similarity information for all pairwise examples in the data. If a label is known for an example, it is placed on the example and 'propagated' or 'diffused' to other examples that have no known labels. This can be seen as running a Markov chain over the graph.

We construct a graph $G = (V, E)$ where the set of vertices V are English and Chinese sentences $(e_1, ..., e_M)$ and $(c_1, ..., c_N)$. There are then $M \times N$ graph edges between the Chinese and English sides, where the edge weights w_{ij} represent the cosine similarity $\cos(e_i, c_j)$. Edges among sentences in the same language are also created to represent the document structure. We set $w_{ij} = 1/dist(c_i, c_j)$ if c_i and c_j are from the same paragraphs, where *dist* is the distance (number of intervening sentences) between c_i and c_j; if they are in different paragraphs, we set $w_{ij} = 0$. The graph allows us to represent global information about all similarity links and document structure. Figure 4 gives a pictorial example.

We initialise the graph by labelling the English sentence to be inserted with label +1, and all other sentences with label 0. The goal is to find a labelling over $(c_1, c_2, ..., c_N)$ by propagating the existing labels. After label propagation, each Chinese sentence will receive a label in the range $[0, 1]$. The position after the Chinese sentence with the maximum value is then chosen to be the place of insertion. Intuitively, Chinese sentences that have a high probability link to the English sentence with +1 label will more likely be the insertion position.

Label propagation can be performed by an iterative Markov chain computation, or by direct eigenvector computation [17]. In the latter case, the following objective can be used:

$$\min_{\mathbf{f}} \sum_{(i,j) \in E} w_{ij} (f_i - f_j)^2 \tag{1}$$

where f_i is the labelling on vertex i, which is capped at +1 or 0 for English sentences and left undetermined for Chinese sentences. $\mathbf{f}$ is an $(N + M)$-dimensional vector of labels. The objective accomplishes label propagation by forcing a pair of vertices $(i,\ j)$ to have similar labels f_i and f_j if the edge weight w_{ij} is large. $\mathbf{f}$ is computed by taking the eigenvectors of the graph Laplacian [17].

3 Experiments

3.1 Data Set and Preprocessing

We collect a set of articles from Wikipedia to evaluate our proposed framework. We first found a set of 2,792 articles that are featured articles in English (as of 17 February 2010)[4]. Featured articles are well-developed and mature articles and they represent good source of new information for other language editions.

From within this set, we performed extensive manual annotation on nine articles on a broad range of topics. These articles contain a total of about 2,000 English sentences and 1,600 Chinese sentences. Two bilingual-speaking annotators work to identify which English sentences contain new information. If an English sentence does *not* provide new information, the annotators label which Chinese sentence it aligns to. Alignments of multiple Chinese sentences to one English sentence (and vice versa) are allowed. Further, when a Chinese sentence only contains partial information, it is also considered as aligned to the English.

[4] `http://en.wikipedia.org/wiki/Wikipedia:Featured_articles`

Table 1. Articles selected for manual inspection and alignment. The table shows the number of sentences in the English and Chinese editions. The ‘aligned’ column shows the number of sentences in English that are aligned to some sentences in Chinese.

Article	Sent. (EN)	Sent. (ZH)	Aligned
Acetic acid	194	169	155
Angkor Wat	149	222	71
Australia	258	229	72
Ayumi Hamasaki	227	306	114
Battle of Cannae	221	149	100
Boeing 747	356	185	298
H II region	116	81	103
India	245	156	67
Knights Templar	156	119	39

The manual annotation is a laborious process since on average the featured articles selected have 210 sentences in one English document and substantial amounts in Chinese. The manual annotation took 2-3 hours on average per article. The inter-annotator agreement was high, with $\kappa = 0.826$, determined on 3 articles (732 sentences) of overlapping annotation.

3.2 Experimental Setup and Results

Identifying New Information. We first present experiments on identifying sentences that contain new information. Our test set contains the nine articles manually annotated. A sentence in the English edition is considered to be containing new information if it is not aligned to any Chinese sentence. We compare four different methods in this classification task:

1. **Heuristic**: Heuristic method that uses only cosine similarity information (Section 2.2)
2. **Self-train**: Linear SVM trained on top/bottom (N=30%) group of sentences (Section 2.2)[5].
3. **Cheat**: Similar to the SVM above, but the true labels are used in training. This is a diagnostic to see to what extent the assumption of self-training holds true.
4. **Random**: Randomly classifying a sentence as containing either new or existing information.

To avoid having to decide on a particular similarity threshold, we evaluate using the area under the precision-recall curve (AUC) for each annotated document. A higher value of AUC in general means that precision is higher for a given recall level. The results are shown in Table 2.

Both heuristic and the self-train SVM achieved high AUC values of 70-95%, for all but two articles, showing that most new information can be captured automatically. Performance on two articles gave surprisingly low AUC scores (‘Boeing 747’ and ‘H II Region’). We discovered that they are the only articles in which the number of aligned pairs is larger than the number of Chinese sentences

[5] We use SVM-rank, a publicly-available SVM tool: `http://svmlight.joachims.org`

Table 2. Results of new information identification. Numbers refer to the area under the precision-recall curve in percentage

Article	Heuristic	Self-train	Cheat	Random
Acetic Acid	70.86	**72.24**	79.69	24.36
Angkor Wat	81.36	**81.77**	86.45	49.85
Australia	92.97	**93.04**	93.16	74.72
Ayumi Hamasaki	**72.59**	71.26	72.32	50.14
Battle of Cannae	84.60	**85.14**	83.14	54.69
Boeing 747	**54.18**	52.95	54.16	19.22
H II Region	54.94	**55.65**	71.30	46.83
India	95.40	**95.63**	95.75	71.24
Knights Templar	**89.38**	88.59	93.63	79.17

Table 3. Percentage of weights assigned to features for different methods

	Heuristic	**Self-train**	**Cheat**
Similarity features	100	52 ± 3	35 ± 12
Neighbour features	0	16 ± 2	39 ± 15
Entropy features	0	32 ± 2	27 ± 12

(see Table 2). This implies that the Chinese sentences are longer, such that multiple English sentences were aligned to one Chinese sentence. As a result similarity between sentences in these two articles tend to be much lower than expected. Overall, we note that for articles in which lengths of sentences are not drastically different across languages, our methods gave reasonable results.

Further, we analysed the SVM models to see what kinds of features are important. We summed up (linearly) the SVM's weights corresponding to each of the three main types of features. Table 3 shows that for the self-trained SVM, similarity features are deemed most useful and account for 52% of the weights, and that entropy features are second. This is expected because all training samples have extreme similarity values. However, for the cheating SVM neighbour features are much more important, accounting for 39% of the weights. This shows that whether neighbouring sentences contain new information is a useful hint on how a sentence should be classified.

In summary, both heuristic and self-trained SVM give satisfactory performances by achieving over 70% AUC in most cases, with the former performing slightly better in some cases. The cheating SVM suggests that we can significantly improve classification results even if only some of the sentences are labelled manually.

Cross-lingual Sentence Insertion. Our second task is sentence insertion. For each article, we randomly select a number of sentences in English that has been manually aligned to some Chinese sentences. We then cover up these alignments, simulating the situation that we have a set of new sentences whose correct positions in the Chinese edition are known. We test the performance of the following four methods:

1. **Manual alignments**: A method based on the manually created alignments of other sentences (Section 2.3).

Table 4. Sentence insertion results, with 30% new information

Method	Average Distance	Section Accuracy	Paragraph Accuracy
Manual alignment	11.5	72.8%	43.1%
Similarity alignment	19.3	57.5%	35.5%
Label prop (para)	**10.5**	**83.9%**	**72.7%**
Label prop (sec)	13.2	81.7%	71.5%

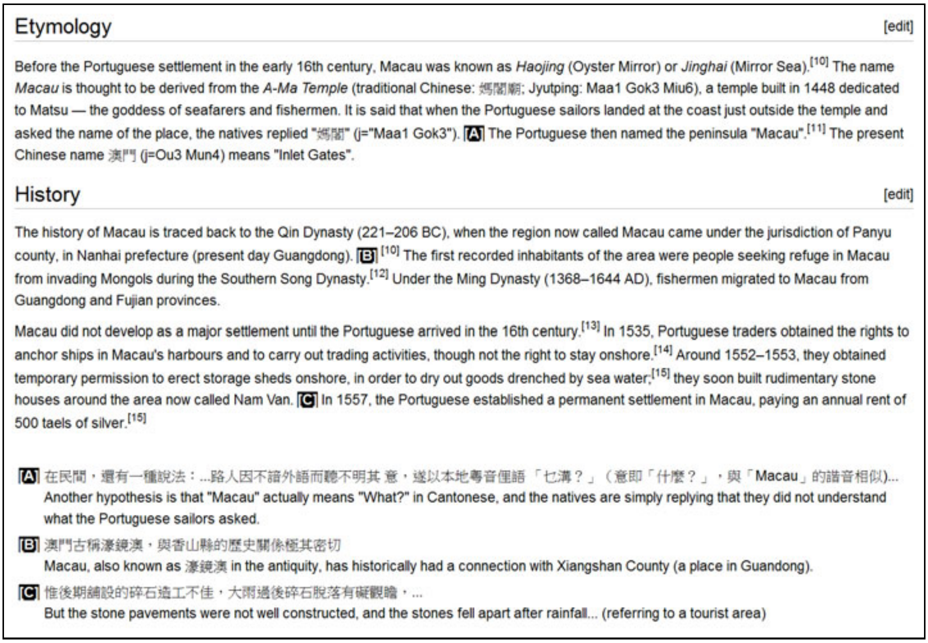
Table 5. Sentence insertion results, with 50% new information

Method	Average Distance	Section Accuracy	Paragraph Accuracy
Manual alignment	14.9	70.7%	39.7%
Similarity alignment	17.6	59.3%	34.3%
Label prop (para)	**11.3**	**82.9%**	**76.8%**
Label prop (sec)	14.0	81.9%	76.4%

2. **Similarity alignments**: A method based on the alignments automatically created by sentence cosine similarity. We choose the top 100 pairs of sentences as correct alignments (Section 2.3).
3. **Paragraph-based label propagation**: The method described in Section 2.3.
4. **Section-based label propagation**: Similar to the above method, but we also create links between sentences appearing in the same section.

To measure performance, we use three different evaluation metrics, averaged over the nine test articles: (1) **Average Distance** (distance in number of sentences between the predicted position and the true insertion position), (2) **Section Accuracy** (whether sentence is inserted into the correct section), and (3) **Paragraph Accuracy** (whether sentence is inserted into the correct paragraph). We decide not to measure accuracy at the sentence level. This is because very often there is no single 'best position' where a sentence should be inserted. Instead, suggesting a paragraph to which a sentence should be inserted is already of great assistance to an editor.

We conduct experiments with different amounts of test data (30% and 50%). The results are in Table 4 and 5. We observe that label propagation outperforms both manual and similarity-based alignments in all three metrics. This is a nice result considering that manual alignment uses true alignments while label propagation does not. The implication is that true alignments are actually not necessary for global graph-based methods. The similarity-based alignment does not use manual information but it performs much worse (e.g. up to 26% decrease in section accuracy). In both tables, we see that different variants of the graph lead to slightly different accuracies for label propagation. It is well-known in the semi-supervised learning literature that optimising the graph structure may lead to better results [16], and we leave that as future work. Here we do not optimise the graph because we want to stay within an unsupervised learning setting where minimal human annotation is required for our methods.



Etymology [edit]

Before the Portuguese settlement in the early 16th century, Macau was known as *Haojing* (Oyster Mirror) or *Jinghai* (Mirror Sea).[10] The name *Macau* is thought to be derived from the *A-Ma Temple* (traditional Chinese: 媽閣廟; Jyutping: Maa1 Gok3 Miu6), a temple built in 1448 dedicated to Matsu — the goddess of seafarers and fishermen. It is said that when the Portuguese sailors landed at the coast just outside the temple and asked the name of the place, the natives replied "媽閣" (j="Maa1 Gok3"). **A** The Portuguese then named the peninsula "Macau".[11] The present Chinese name 澳門 (j=Ou3 Mun4) means "Inlet Gates".

History [edit]

The history of Macau is traced back to the Qin Dynasty (221–206 BC), when the region now called Macau came under the jurisdiction of Panyu county, in Nanhai prefecture (present day Guangdong). **B** [10] The first recorded inhabitants of the area were people seeking refuge in Macau from invading Mongols during the Southern Song Dynasty.[12] Under the Ming Dynasty (1368–1644 AD), fishermen migrated to Macau from Guangdong and Fujian provinces.

Macau did not develop as a major settlement until the Portuguese arrived in the 16th century.[13] In 1535, Portuguese traders obtained the rights to anchor ships in Macau's harbours and to carry out trading activities, though not the right to stay onshore.[14] Around 1552–1553, they obtained temporary permission to erect storage sheds onshore, in order to dry out goods drenched by sea water;[15] they soon built rudimentary stone houses around the area now called Nam Van. **C** In 1557, the Portuguese established a permanent settlement in Macau, paying an annual rent of 500 taels of silver.[15]

A 在民間，還有一種說法：...路人因不諳外語而聽不明其 意，遂以本地粵音俚語「乜溝？」（意即「什麼？」，與「Macau」的諧音相似)...
Another hypothesis is that "Macau" actually means "What?" in Cantonese, and the natives are simply replying that they did not understand what the Portuguese sailors asked.

B 澳門古稱濠鏡澳，與香山縣的歷史關係極其密切
Macau, also known as 濠鏡澳 in the antiquity, has historically had a connection with Xiangshan County (a place in Guandong).

C 惟後期舖設的碎石造工不佳，大雨過後碎石脫落有礙觀瞻，...
But the stone pavements were not well constructed, and the stones fell apart after rainfall... (referring to a tourist area)

Fig. 5. Sentence insertion for the article ‘Macau’. We show part of the English edition and three sentences containing new information in Chinese. The alphabets indicate the suggested insertion positions.

4 Discussion

4.1 An Example of System Output

We apply our framework to the article ‘Macau’ to identify new information in Chinese and insert sentences into the English edition[6]. The article is a featured article in Chinese but not in English. Figure 5 shows three sentences identified as containing new information using the self-trained SVM approach, and the positions where our label propagation algorithm suggests they should be inserted.

Sentence (A) and (B) are correct insertions. The former provides information about the origin of the name of Macau, while the latter provides information about Macau’s geographical relations. However, while being a sentence containing new information for the English edition, Sentence (C) is an example of incorrect insertion. The sentence is irrelevant to the paragraph, but is inserted at that position because of the word ‘stone’, which is a rare word throughout the documents. Since this sentence refers to a topic that is not present in the English edition, it becomes difficult for the algorithm to find a correct position. Overall, our algorithm works well when sentences contain information that is new but is still related to the topics present in the target document.

[6] This is the opposite direction of the experiments, to demonstrate the flexibility of our approach.

4.2 Other Issues and Limitations

Several issues deserve further attention. Firstly, we treat sentences as the basic units of information, which has resulted in certain limitations. We plan to study how these can be solved. Nevertheless, the existence of this problem actually motivate our work, because this means that it requires even more effort from human editors to distinguish between new and existing information.

Secondly, our current similarity measure only takes into account lexical similarity and may overlook synonyms. The accuracy of similarity also depends on the result of machine translation. While Google Translate mostly return sufficiently good translations for measuring sentences similarity on the lexical level, to further improve performance we will consider incorporating the translation model into our framework instead of treating it as a black box.

Finally, in this work we do not consider the 'value' of the sentences. As articles are constantly under revision, sentences may be deleted for various reasons. Vandalism is also not uncommon in Wikipedia. Hence, it would be desirable to determine whether a sentence (and the information it contains) is valuable to be inserted into other languages. We can incorporate into our framework methods for vandalism detection [15], or methods for assessing the credibility of the editors who wrote the sentences [7].

5 Related Works

While there are no directly comparable works, some authors have studied related problems. For example, Chen et al. [5] propose a method based on sentence features for inserting new information into existing texts in a monolingual setting. Our work differs from theirs in that we consider a cross-lingual setting and can therefore take advantage of the document structures of both the source and target article. Our method also does not require supervised labels.

Adar et al. [2] introduce an automated system called Ziggurat for aligning and complementing infoboxes across different languages in Wikipedia. Tacchini et al. [11] presents some experiments on data fusion across languages using the DBpedia framework. Our proposed framework can be used to handle the texts of the articles and is therefore more general and applicable to other settings.

Sauper and Barzilay [9] propose a method for generating Wikipedia articles by inducing an article template automatically and retrieving relevant texts from the Web. We believe that this method would be complementary to our proposal, because our method relies on the fact that the articles already contain some information. In cases when a topic simply does not exist, an automatically generated article will be a very good starting point for cross-lingual enrichment.

Lapata [8] proposes using a Markov chain to model the structure of a document. Barzilay and Elhadad [4] proposes a method for sentence alignment that involves first matching larger text fragments by clustering and further refine these matches by local similarity. These techniques, however, require a large corpus for training, while our proposed model operates only on the article level and does not require any labels.

Finally, our work is also related to the task of automatic extraction of parallel sentences from comparable corpora [10], as sentences that are not found to have any correspondence in another language should contain new information. We plan to investigate how methods for this task can be incorporated into our framework to improve performance.

6 Conclusions

We propose a new task, 'cross-lingual document enrichment', of which the goal is to assist editors in bridging the information gap within multi-lingual document collections. Our contributions include (1) a framework for addressing this task in terms of two sub-tasks: new information identification and cross-lingual sentence insertion; and (2) a proof-of-concept system using a novel combination of NLP and machine learning techniques. While there are other ways for improvement, our system already demonstrates the ability to significantly alleviate the load for human editors. In addition to investigating the issues mentioned in Section 4.2, we will also carry out evaluations of larger scale on various datasets. We believe that this is a promising research direction for NLP to impact the creation and maintenance of vast multi-lingual document collections.

References

1. Abney, S.: Bootstrapping. In: 40th Annual Meeting of the Association for Computational Linguistics (2002)
2. Adar, E., Skinner, M., Weld, D.S.: Information arbitrage across multi-lingual Wikipedia. In: WSDM 2009, pp. 94–103 (2009)
3. Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Cortes, C., Mohri, M.: Polynomial Semantic Indexing. In: NIPS (2009)
4. Barzilay, R., Elhadad, N.: Sentence alignment for monolingual comparable corpora. In: EMNLP 2003, pp. 25–32 (2003)
5. Chen, E., Snyder, B., Barzilay, R.: Incremental text structuring with online hierarchical ranking. In: EMNLP-CoNLL 2007, pp. 83–91 (2007)
6. Hecht, B., Gergle, D.: The Tower of Babel meets Web 2.0: user-generated content and its applications in a multilingual context. In: CHI 2010, pp. 291–300 (2010)
7. Javanmardi, S., Lopes, C., Baldi, P.: Modeling user reputation in Wikipedia. Journal of Statistical Analysis and Data Mining 3(2), 126–139 (2010)
8. Lapata, M.: Probabilistic text structuring: experiments with sentence ordering. In: ACL 2003, Morristown, NJ, USA, pp. 545–552 (2003)
9. Sauper, C., Barzilay, R.: Automatically generating Wikipedia articles: A structure-aware approach. In: ACL 2009, pp. 208–216 (2009)
10. Smith, J., Quirk, C., Toutanova, K.: Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In: ACL 2010 (2010)
11. Tacchini, E., Schultz, A., Bizer, C.: Experiments with Wikipedia cross-language data fusion. In: Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web, ESWC 2009 (2009)
12. Wikipedia. Translation of the week (2010), `http://meta.wikimedia.org/wiki/Translation_of_the_week` (accessed May 10, 2010)

13. Wikipedia. Wikipedia machine translation project (2010), http://meta.wikimedia.org/wiki/Wikipedia_Machine_Translation_Project (accessed May 10, 2010)
14. Wikipedia. Wikipedia:translate (2010), http://en.wikipedia.org/wiki/Wikipedia:Translation (accessed May 10, 2010)
15. Wang, W., McKeown, K.: Got You!: Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling. In: COLING 2010, pp. 1146–1154 (2010)
16. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison (2005), http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
17. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of International Conference on Machine Learning (2003)