Managing Information Disparity in Multi-lingual Document Collections

KEVIN DUH, NTT Communication Science Laboratories
CHING-MAN AU YEUNG, NTT Communication Science Laboratories
TOMOHARU IWATA, NTT Communication Science Laboratories
MASAAKI NAGATA, NTT Communication Science Laboratories

Information disparity is a major challenge with multi-lingual document collections. When documents are dynamically updated in a distributed fashion, information content among different language editions may gradually diverge. We propose a framework for assisting human editors to manage this information disparity, using tools from machine translation and machine learning. Given source and target documents in two different languages, our system automatically identifies information nuggets that are new with respect to target and suggests positions to place their translations. We perform both real-world experiments and large-scale simulations on Wikipedia documents and conclude our system is effective in a variety of scenarios.

Categories and Subject Descriptors: H.5.3 [**Group and Organization Interfaces**]: Web-based interaction; I.2.7 [**Natural Language Processing**]: Text analysis; I.7.1 [**Document and Text Editing**]: Languages

General Terms: Algorithms, Languages, Experimentation

Additional Key Words and Phrases: Cross-Lingual Methods, Document Management Systems, Machine Translation Applications

ACM Reference Format:
Kevin Duh, Ching-man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing Information Disparity in Multi-lingual Document Collections. *ACM Trans. Speech Lang. Process.* 9, 4, Article 39 (March 2010), 29 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Multi-lingual document collections have become important resources in this global age. In the past, document collections were often constructed with monolingual audiences in mind. Nowadays, information needs to be spread to multiple language communities very quickly, making the creation and maintenance of multi-lingual document collections an important topic.

Scenarios of this kind are abundant: International organizations have to maintain documents in different languages to be consumed by members from different countries. Multinational corporations face similar situations when they need to localize guidelines and product specifications in different places all over the world. In addition, the rising popularity of distributed collaboration on the World Wide Web has resulted in the development of multi-lingual information repositories such as Wikipedia[1] and Wik-

[1]Wikipedia: http://www.wikipedia.org/

Author's addresses: K. Duh, (Current address) Graduate School of Information Science, Nara Institute of Science and Technology; C.-M. Au Yeung, (Current address) Noah's Ark Lab, Huawei; T. Iwata and M. Nagata, NTT Communication Science Laboratories, NTT Corporation.

DOI: http://dx.doi.org/10.1145/0000000.0000000

iTravel[2], which consist of articles with multiple editions of different languages; editors may wish to improve native articles using other language editions as reference.

One major challenge in managing multi-lingual document collections is that information in these documents may be continuously updated. This leads to possible *information disparity* among different language editions. Information disparity is especially problematic for document collections created in a distributed fashion, such as Wikipedia, where different authors may independently update different language versions. The different language versions may not be intended as exact translations, but instead have variations in localized content and document structure. In this case, translators in charge of reducing information disparity are burdened with additional work besides actual translation, such as deciding exactly what piece of information needs to be translated and identifying where to insert the result in the target document. This is an inefficient use of human translator time and is likely to cause delays in having the most up-to-date information appear in all languages.

For instance, consider Wikipedia, a collaboratively edited encyclopedia encompassing over 250 languages. Despite its multilinguality, there are significant differences among language editions in terms of size and quality [Hecht and Gergle 2010]. While various projects[3] have attempted to bridge the information disparity, the focus has been on translating existing articles in their entirety. Few projects focus on maintaining and synchronizing along language versions as articles are updated continuously, because too much human effort is required.

In view of this problem, we propose a framework, termed **cross-lingual document enrichment**, for managing information disparity using tools from machine translation and machine learning. Given two documents in different languages, our system first uses a MT-based cross-lingual similarity metric to identify sentences that contribute to information disparity. Then, we employ a graph-based method to predict the best position to insert the translation in the target document structure. The benefit of such a system is that it can greatly reduce the effort required to manage a multilingual document collection: the human translator can focus on the actual translation work while our system provides suggestions for what to translate and where to insert the result.[4]

The contribution of this paper is two-fold:

(1) Firstly, we propose cross-lingual document enrichment as a novel research problem (Section 2) and provide automatic unsupervised solutions for managing information disparity (Sections 3 and 4). As far as we know, only a few previous works address the information disparity challenge in multi-lingual collections (Section 7).
(2) Secondly, we perform two comprehensive evaluations, one using realistic data and another involving large-scale simulation. On the realistic data, our system demonstrates its effectiveness in bridging information disparity inherent in Wikipedia (Section 5). On the large-scale simulated data created from machine translation bitext, we explore in depth how our system performs under a variety of conditions (Section 6).[5]

[2]WikiTravel: http://wikitravel.org

[3]e.g. Translation of the Week: http://meta.wikimedia.org/wiki/Translation_of_the_week; Wikipedia Machine Translation Project: http://meta.wikimedia.org/wiki/Machine_Translation_Project

[4]One may expect that a more comprehensive system would do away with human translations and automatically synchronize content using MT. We do not consider this here because such a task would require an MT system of very high quality. Our goal is to assist the management of information disparity, not the actual process of translation.

[5]This large-scale study is a new contribution compared to our previous work [Au Yeung et al. 2011].

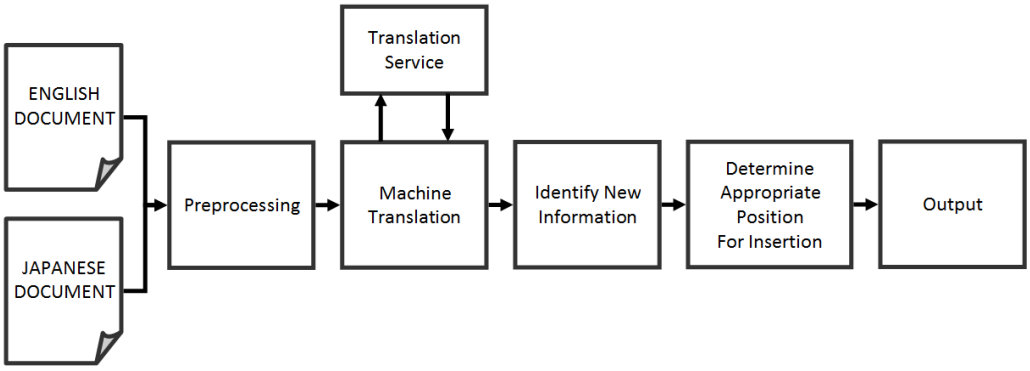


Fig. 1. The system design of our framework for cross-lingual document enrichment. Machine translation is used to map the two language editions into the same language such that similarity between sentences can be computed. Based on the similarity the system then identifies sentences containing new information and subsequently suggests appropriate positions for insertion.

2. GENERAL FRAMEWORK: CROSS-LINGUAL DOCUMENT ENRICHMENT

2.1. System Overview

In this section we present an overview of our proposed framework for cross-lingual document enrichment. While the framework is independent of the languages involved, for concreteness we will assume that we are dealing with information disparity in a collection that contains English and Japanese documents. We focus on cases where the Japanese documents are more up-to-date and contain more information than the English documents. As a result, our task is to assist the task of enriching English documents with new information found in their Japanese counterparts.

Our framework makes two general assumptions. First, we assume the enrichment process is directional, i.e. using Japanese documents to enrich English. There may be situations where bidirectional, mutual enrichment is desired, but this increases the complexity of the problem. Directional enrichment can be very suitable in cases where the editor is mainly interested in improving the documents in his or her own native language, while using references from any source language. Second, we treat *sentences* as the basic units of information. One may argue that information granularity may cross sentence boundaries, but taking that into account also increases system complexity. Our focus is to assist human editors to manage information disparity (and not to build a fully-automated system at this time); we believe these two assumptions are reasonable for this this application scenario.

Figure 1 depicts the overall system design of our framework. For each article, English and Japanese documents are preprocessed to remove formatting information. Sentences are extracted and labeled by section and paragraph IDs. We then use a machine translation system to translate all Japanese sentences to English. In practice we can use a variety of ways to map sentences from the two sides to the same symbol set. The goal is to enable sentence similarity computation between two languages; Section 2.2 discusses the details.

We are then ready to perform the following two tasks:

— **(Task 1) New information identification**: Given two sets of sentences (one from the source Japanese document and another from the target English document), identify a subset (of Japanese sentences) that contains information not found in English. (Section 3).

— **(Task 2) Cross-lingual sentence insertion**: Given a set of sentences obtained from the above task, determine for each of them a suitable location for insertion in the target English document. (Section 4).

The output of the system will be a set of sentences that contain new information that is not present in the target document, and a set of appropriate positions in the target document where these sentences should be inserted. An editor of the document collection can then determine whether these sentences (or the information they contain) are suitable for the target document, and translate them either by referring to the machine translated sentence or by obtaining a new translation from a human translator.

To clarify the scope of this work, note that we can classify a multi-lingual collection along two axes: First, is the collection static after creation or dynamically updated continuously? Second, is the content and structure meant to be exact translations (i.e. parallel) or only meant to be comparable (i.e. carry some amount of shared information but allowing for some divergence). Table I shows some examples of each. Our focus here is on dynamically-updated and comparable collections, as this category poses the most challenge from the information management perspective. This category is also the most pertinent for document collections that arise due to distributed collaboration on the World Wide Web, which is of much importance. (Note that Task 1 is trivial for statically-created collections, while Task 2 is trivial for parallel collections.)

Table I. Categorization of multi-lingual document collections and some examples. In practice, the division may not be clear-cut as shown here and there may be some examples that fall under multiple categories. Our focus is dynamic and comparable collections.

	Parallel Content/Structure	Comparable Content/Structure
Static Creation	Parliament proceedings	Multilingual newspapers
Dynamic Update	Technical FAQs	Wikipedia, Product localization

2.2. Measuring Cross-Lingual Sentence Similarity

A critical element in our system is the similarity metric between sentences of different languages. The reliability of this metric influences the results of both Task 1 and Task 2. In this section, we describe how we measure cross-lingual sentence similarity using machine translation (MT).

2.2.1. MT-based Similarity Metric. We use MT to map two sentences from different languages into the same symbol set, so that conventional mono-lingual similarity metrics can be applied. While we translate from Japanese to English, note that we can as well as translate the English to Japanese, translate both editions to French, or any combination of the above methods. 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 such as LSA [Deerwester et al. 1990] or PSA [Bai et al. 2009]. We can also translate using bilingual dictionaries rather than full-scale MT [Rapp 1999], which may be extracted from less stringent resources such as comparable corpora. The goal is to just convert different languages into a comparable representation.

After we translate all Japanese sentences in a document to English, we employ a straightforward bag-of-words approach to characterize the sentences. Each English sentence e is represented by its unigram term vector. Each Japanese sentence j is represented by the term vector computed from its English translation. For a document pair, we extract a vocabulary of size V after stop-word removal and stemming; the vectors e and j are sparse V-dimensional term vectors, where terms are weighted by the TF-IDF scheme, i.e. the term element of vector e equals to the number of times that term occurs in sentence e, divided by the number of sentences in the document

pair where the term occurs.[6] We then use standard cosine similarity: given e and j, the similarity is defined as

$$\cos(e, j) = \frac{e^T \cdot j}{||e|| \cdot ||j||} \quad (1)$$

where the numerator is the dot product of the two vectors and denominator is normalized by the L2-norms of each vector [Manning et al. 2008].

We consider cosine similarity because it is one of the most basic approaches, fast to compute on large collections, and requires no additional resources. One may also consider other resource-lean metrics such as Jaccard or Dice, or metrics enhanced with semantic knowledge, e.g. [Budanitsky and Hirst 2006]. The reliability of this similarity metric depends on the quality of the MT output; we will also evaluate this impact in the experiments.

2.2.2. Using N-Best Translation Candidates. If the machine translation system outputs a set of n-best translation candidates for a given sentence, we can take advantage of the alternative translations to improve the similarity metric. This is because the top translation given by an MT system may not necessary be the most appropriate translation in practice. N-best lists could perhaps contain synonyms, which would increase the reliability of our similarity metric. Let the N-best list of a Japanese translation be $\{j^{(c)}\}_{c=1,2,\ldots,N}$, where $j^{(1)}$ is the most confident translation and $j^{(N)}$ is the N-th confident translation. We consider a few ways of utilizing the n-best list to improve similarity calculation.

(1) **1best**: The baseline is to simply use the first result in the n-best list: $S_{1best} = \cos(e, j^{(1)})$
(2) **nbest-prob**: Statistical MT systems usually provide a confidence value or likelihood score for each candidate in the N-best list. One way to integrate information from multiple candidates is a weighted combination of each candidate's cosine similarity based on these values. Here we normalize the likelihood scores over the N-best list in order to obtain $p(j^{(c)})$, the probability of candidate $j^{(c)}$. Then we compute: $S_{prob} = \sum_{c=1}^{N} p(j^{(c)}) \cdot \cos(e, j^{(c)})$
(3) **nbest-concat**: An alternative approach to integrating N-best information into the cross-lingual metric is to concatenate all Japanese candidates into a single sentence, then compute cosine. This is equivalent to accumulating the term frequency over all candidates, thereby increasing the potential coverage of our bag-of-words representation. The increased length in the concatenated sentence and the ordering of the words in the new sentence is not important because cosine measure is invariant to those changes: $S_{cat} = \cos(e, J)$ where J is the concatenation of all candidates.
(4) **nbest-oracle**: Ideally, it would be good to be able to determine which candidate in the n-best list is the best translation. Assuming we have the correct reference of a translation, we can calculate the similarity between this reference and all the candidates in the n-best list. The candidate that achieves the highest similarity can be considered as the best candidate, and can be used in subsequent tasks. While we do not have references in practice, we study the performance of this method in our experiments to investigate the effects of translation quality on the performance of

[6] In contrast to conventional TF-IDF which is applied to documents as units, we are operating with sentences as units. So TF (term frequency) is counted within the sentence and IDF (inverse document frequency) is actually inverse sentence frequency.

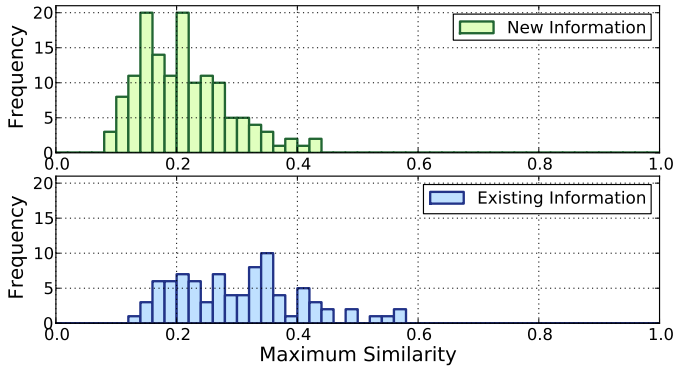


Fig. 2. Distribution of maximum similarity values of sentences with new or existing information in a sample document.

our proposed algorithms: $S_{oracle} = \cos(e, j^o)$, where $o = \arg\max_n cos(j^r, j^{(c)})$ and j^r is the reference translation.

2.2.3. Alternative Similarity Metrics. While our system employs MT and bag-of-words cosine as the cross-lingual similarity metric, other metrics could be plugged in as well. For example, rather than a bag-of-words metric, one could employ more sophisticated semantic inference engines from the textual entailment field [Mehdad et al. 2010]. Another approach is to do away with MT altogether: recent Bayesian techniques such as polylingual topic models [Mimno et al. 2009] can directly estimate topic similarity using comparable (not parallel) multi-lingual corpora. Our experiments will examine some of these alternatives and we defer the detailed explanation to relevant experiment section (Section 6.1.3).

3. TASK 1: NEW INFORMATION IDENTIFICATION

Our first task is to identify Japanese sentences containing information that is not present in the English edition. We first describe an unsupervised method that makes use of only the cross-lingual similarity scores. In addition, we also consider a supervised method that takes advantage of partially-labeled alignments between sentences if available.

The task is formally defined as follows: Given a document pair with M Japanese sentences $\{j_m\}_{m=1,\ldots,M}$ and N English sentences, $\{e_n\}_{n=1,\ldots,N}$, find the subset of Japanese sentences within $\{m = 1, \ldots, M\}$ such that they are considered new information with respect to the English.

3.1. The MaxSim Method

Intuitively, a new Japanese sentence should have low similarity to all of the existing English sentences. On the other hand, a Japanese sentence that contains existing information should have high similarity to at least one English sentence. As a result, the *maximum* similarity of a Japanese sentence to any English sentence can be a good predictor of whether the sentence itself contains new information.

This gives a straightforward algorithm, MaxSim, shown in Algorithm 1. First, we compute the pair-wise cross-lingual similarity between Japanese sentences and English sentences, then obtain the maximum similarity of each Japanese sentence. The Japanese sentences are then sorted by their MaxSim value in ascending order and returned by the algorithm as a ranked list. The human editor can then check this list from the top, which are likely to contain new information. We can alternatively set a

threshold on the MaxSim value needed to be returned, using estimation techniques from the novelty detection field [Markou and Singh 2003].

Algorithm 1 MaxSim algorithm for new information identification

Input: Two set of sentences $\{j_m\}_{m=1,...,M}$, $\{e_n\}_{n=1,...,N}$
Output: A subset or ranking of $\{j_m\}$ likely to contain new information

1: **for** $m = 1, ..., M$ **do**
2: **for** $n = 1, ..., N$ **do**
3: compute cross-lingual similarity $S(e_m, j_n)$ (Section 2.2.2)
4: **end for**
5: maxsim(j_m) = $\max_n S(e_m, j_n)$
6: **end for**
7: Return sentences j_m ranked by increasing maxsim(j_m)
8: Alternatively, return j_m whose maxsim(j_m) is smaller than a threshold

Figure 2 shows the distribution of maximum cosine similarity values for sentences that contain new information and those that contain existing information in a sample article. It is interesting to note the asymmetry: for new sentences, maxsim value is always low (rarely greater than 0.4); for sentences containing existing information, maxsim value may exhibit a larger range. The reason: cosine similarity may not be high even for existing information because incorrect translations or lack of true semantic matching limit the overlap of words. On the other hand, we can be quite sure that our MaxSim value reliably filters out a portion of existing information, since high MaxSim value is a clear indicator of information overlap. In our experiments we will see that such straightforward method actually gives relatively good results.

3.2. A Classifier approach using Partial Labels

While MaxSim is an unsupervised method, now we discuss machine learning alternatives for cases when partial labels are available. There might be situations where cross-lingual sentence alignments are available in small amounts, and these are invaluable for improving the system performance. For example, documents in different languages might be created at the same time in the past, and sentences in different languages are direct translations of those in a master document. While new content may be added to different language editions separately at a later time, alignments between sentences that were written in the very beginning can be useful in identifying information disparity at a later time.

If partial labels are available, we can setup a classification task as follows: given an article with Japanese sentences $(j_1, j_2, ..., j_M)$, label each sentence j_i with $\{+1, -1\}$ where $+1$ indicates that the sentence contains new information and -1 indicates otherwise. The remaining Japanese sentences, where labels are not given, become the test samples. We can introduce several features and train a classifier for identifying which of the remaining sentences are new information.

A feature vector is defined for each sentence j_m. The main types of features are:

— **MaxSim and variants** (5 features): Maximum cosine similarity of j_m, i.e. $\max_n \cos(e_n, j_m)$. This is the feature used in the MaxSim method. We also include variants in the form of top-k averages, $\frac{1}{k}\sum_{k\in K}\cos(e_n, j_m)$, where K is the set of k pairs with the highest cosine similarities ($k = 2, ..5$). The higher these values, the more likely one is existing information.
— **Minimum similarity** (1 feature): The minimum similarity value $\min_n \cos(e_n, j_m)$ per sentence is included to act as a calibration.

— **Neighbors** (2 features): Maximum cosine similarity of the neighbors, j_{m+1} and j_{m-1}. The idea is that if the neighbors have low similarity, then more likely j_m will contain new information, and the opposite is also likely to be true.
— **Entropy** (1 feature): Entropy of similarity values of j_m, where similarity distribution is converted into probability distribution by:

$$-\sum_n \frac{\cos(e_n, j_m)}{\sum_{n'} \cos(e_{n'}, j_m)} \log(\frac{\cos(e_n, j_m)}{\sum_{n'} \cos(e_{n'}, j_m)}) \quad (2)$$

This feature counteracts situations where particular words lead to high cosine values for all sentences. Intuitively, if a Japanese sentence contains existing information, it should only be matched to a small number of English sentences, and would achieve low entropy.

For each of the above nine features, we also compute the deviation from its average of all samples j_m in the document, e.g. the MaxSim deviation feature for j_m would be: $\max_n \cos(e_n, j_m) - \frac{1}{m}\sum_m \max_n \cos(e_n, j_m)$, giving a total of 18 features. We train our classifier using a fast linear SVM classifier [Joachims 2006]. We choose a fast training algorithm because we train a different classifier for each document pair that contains partial labels, thus eliminating the worry of domain differences. For instance, MaxSim values may have different ranges for different documents because the quality of MT varies based on domain. In our experiments, we will see how this additional partial label, when available, can be used to improve upon MaxSim.

4. TASK 2: CROSS-LINGUAL SENTENCE INSERTION

Given the sentences identified in the previous task, we now focus on how we can determine the most appropriate positions in the target document where these sentences can be inserted. We formulate the task as: given a Japanese sentence j_m, finding a sentence e_n in the English edition after which (the translation of) the new sentence should be inserted. We consider two methods to solve this problem.

4.1. Heuristic Insertion

Intuitively, the sentence should be inserted in a way that maintains the order of discourse or the flow of the article. Thus, a reasonable scheme is as follows. We look for a Japanese sentence before j_m, say j_{m-1} that is aligned to an English sentence e_k. By "aligned", we mean that j_{m-1} and e_k are determined to have equivalent information. Since e_k corresponds to j_{m-1}, it becomes natural that j_m when translated into English should follow e_k. If j_{m-1} has no corresponding sentence in the English edition, we can repeat the process and check j_{m-2} and so on. Figure 4(a) illustrates this idea.

Now, the above insertion heuristic is implementable when some sentences in Japanese have been aligned to English manually, which is similar to the case of the partial labels in Section 3.2. On the other hand, when there are no alignments (or when the number of alignments is not sufficient relative to the size of the article), we propose to automatically generate the likely alignments. Specifically, we can generate some alignments automatically by selecting pairs of sentences that achieve high values of cosine similarity. Although these alignments are not necessarily correct, they do provide a basis for us to apply the heuristic described above to search for a possible position.

Algorithm 2 shows the complete method. First, we automatically generate an alignment from Japanese j_m to English e_n if the cosine score is above a threshold τ and highest among all j_m alignments (line 4-5). Then, for the test sentence j_t at Japanese position t, we gradually walk up the previous Japanese positions until we find one with an alignment (lines 9-13). We simply return the alignment $A[t'] \in \{1, \ldots, N\}$ as the in-

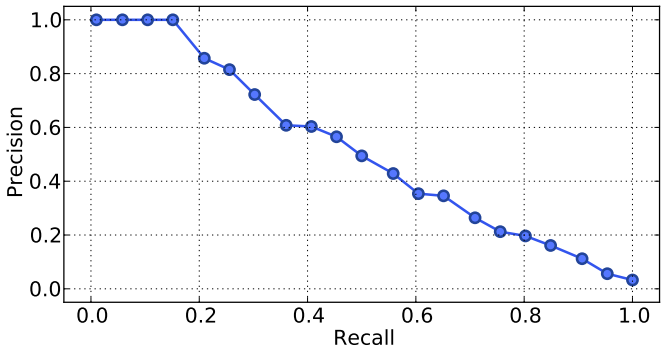


Fig. 3. Precision-recall curve of the similarity-based alignment for a sample article. Pairs of sentences are ordered in descending order of similarity, and precision/recall is evaluated on manual annotation (described in Section 5.1).

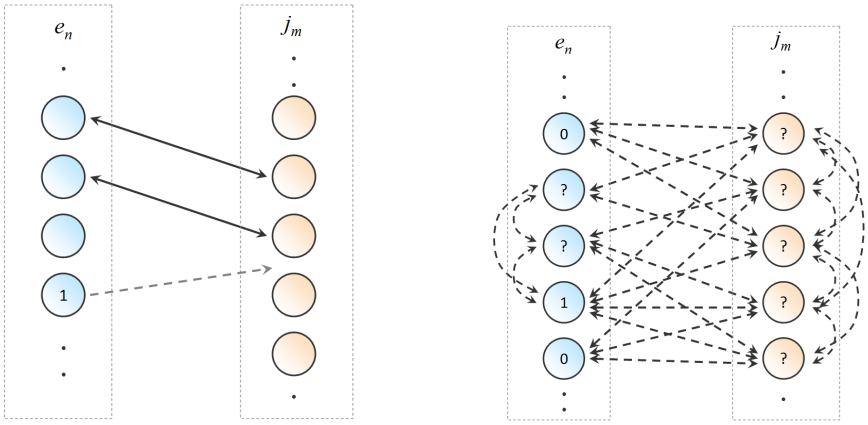
sertion position. In practice, we set the threshold τ to a value that picks up $0.5 * N$ alignments. This leads to a relatively conservative (high) threshold, as the number of pairs are $N * M$. Figure 3 shows the precision-recall curve on a sample document, generated by varying the threshold for determining new vs. existing information and evaluated on manual annotations. Similar to what we see in Section 3, we see that the pairs with high similarity are quite accurate alignments, and precision is perfect in the 0 to 0.2 recall range.

Algorithm 2 Heuristic insertion algorithm for sentence j_t at position t

Input: Two set of sentences $\{j_m\}_{m=1,...,M}$, $\{e_n\}_{n=1,...,N}$
Input: Pair-wise cross-lingual similarity values $S(e_n, j_m)$ for all pairs (n, m)
Output: Target insertion position in $\{n = 1, \ldots, N\}$ of sentence j_t.

Initialize empty hash $A[\cdot]$=undefined, and $B[\cdot]$=0.
for $m = 1, ..., M$ **do**
  for $n = 1, ..., N$ **do**
    if $S(e_n, j_m) > \tau$ and $S(e_n, j_m) > B[m]$ **then**
      Create alignment $A[m] = n$ between e_n and j_m. Set $B[m] = S(e_n, j_m)$.
    end if
  end for
end for
for $(t' = t; t' > 0; t' = t' - 1)$ **do**
  if $A[t']$ is defined **then**
    Return $A[t']$ as insertion point
  end if
end for

The limitation of this relatively simple method, of course, is that we do not have all correct sentence alignments and thus sentences may be inserted into somewhere far away from the correct positions. In addition, highly similar sentences might be concentrated in a particular part of the article. For example, usually the introductory sections in Japanese and English might have more sentences and words in common than the rest of the documents, simply because editors of different languages might



(a) Heuristic Insertion (b) Graph-based Insertion

Fig. 4. Illustration of Insertion methods

choose to focus on different sections thereafter. This skewed distribution may then greatly affect the insertion task.

4.2. Graph-based Method

In view of the limitations of the above methods, we propose a method that is based on graph-based methods (specifically, label propagation [Zhu et al. 2003]). First, we construct an undirected graph $G = (V, E)$ where the set of vertices V are Japanese and English sentences $(j_1, ..., j_M)$ and $(e_1, ..., e_N)$. There are then $M \times N$ graph edges between the Japanese and English sides, where the edge weights w_{nm} represent cross-lingual similarity scores. In addition, edges among sentences in the same language are also created to represent the document structure. We set $w_{nn'} = 1/dist(e_n, e_{n'})$ if e_n and $e_{n'}$ are from the same sections, where $dist$ is the distance (number of intervening sentences) between e_n and $e_{n'}$; if they are in different sections, we set $w_{nn'} = 0$. Edge weights $w_{mm'}$ on the Japanese side is computed analogously. When we talk about any of the above cross-lingual and mono-lingual edges, we use the general notation w_{xy}. The graph allows us to represent global information about all similarity links and document structure. Figure 4(b) gives a pictorial example.

To initialize the graph, we label the Japanese sentence to be inserted into the English edition with label $+1$, and Japanese sentences from other sections with label 0. The goal is to find a labeling over $(e_1, e_2, ..., e_N)$ by propagating the existing labels. After label propagation, each English sentence will receive a label in the range $[0, 1]$. The position after the English sentence with the maximum value is then chosen to be the place of insertion. The intuition is that such an English sentence would contain the most relevant information to that contained in the Japanese sentence to be inserted.

To make this concrete, let us consider Figure 4(b), where some vertices in the source side on the left have been initialized. At each iteration, we "propagate" the labels to the uninitialized nodes along edges that have high weights. Each uninitialized node gets a value depending on the weighted sum of labels from its incoming edges. After many iterations, the labels will converge to some real number between $[0, 1]$, and the node with the highest value is most probable point of insertion.

The above iterative Markov chain interpretation can be implemented by a direct eigenvector computation [Zhu et al. 2003]. We opt for the the eigenvector, rather than

iterative computation, since it is very fast in the case when the graph is not large (which is true in our case since an article pair generally only has hundreds of sentences). The iterative solution at convergence is equivalent to the solution of the following objective:

$$\min_{\mathbf{f}} \sum_{(x,y)\in E} w_{xy}(f_x - f_y)^2 \tag{3}$$

where f_x and f_y are labelings on vertices and the collection of all $N + M$ labelings is represented by vector $\mathbf{f}$. The objective can be minimized by forcing a pair of vertices (x, y) to have similar labels f_x and f_y if the edge weight w_{xy} is large. Specifically, an element/vertex of $\mathbf{f}$ is set to $+1$ in the position of the Japanese sentence that is the new information to be inserted in the English document; it is set to 0 in Japanese sentences far-away, i.e. those from different sections. Let's call this labeled portion of the vector $\mathbf{f_l}$. The goal is to find a labeling for the remaining sentences, which we indicate by the sub-vector $\mathbf{f_u}$. Let us now organize the matrix of edge weights such that $\mathbf{W_{ll}}$ represents all weights within the labeled portion, $\mathbf{W_{uu}}$ represent weights in the unlabeled portion, and $\mathbf{W_{ul}}$ represent weights connecting the two (i.e. many of the cross-lingual similarity values). Then Equation 3 can be solved by the following matrix operation (see [Zhu et al. 2003] for derivation).:

$$\mathbf{f_u} = (\mathbf{D_{uu}} - \mathbf{W_{uu}})^{-1}\mathbf{W_{ul}}\mathbf{f_l} \tag{4}$$

where $\mathbf{D_{uu}}$ is diagonal matrix with elements $d_{xx} = \sum_y w_{xy}$ and the term $\mathbf{D_{uu}} - \mathbf{W_{uu}}$ is called the graph Laplacian. Finally, we find the English element in $\mathbf{f_u}$ that has the highest value and propose it as an insertion point to the human editor. Intuitively, positions with high cross-lingual similarity to the Japanese sentence in question will have high f values; the position with the highest value in practice will also depend on joint interactions with within-document similarities.

5. EXPERIMENTS ON REAL-WORLD WIKIPEDIA DATA

In our first experiment, we crawl Wikipedia for real-world examples of information disparity. We manually annotate the crawled dataset and evaluate how our system performs in realistic scenarios. This section demonstrates a proof-of-concept of our proposed system.

5.1. Data Preparation

We collect and label manually a set of articles from Wikipedia in order to evaluate our proposed framework. First, we found a set of 2,792 articles that are featured articles in English (as of 17 February 2010).[7] Featured articles are well-developed, mature and comprehensive articles, which represent good source of new information for editions in other languages. Our task is to find the new information and insert it in the corresponding Chinese edition.[8]

From within this set, we performed extensive manual annotation on nine articles on a broad range of topics. To focus on a challenging task, we restricted our annotation to article pairs where the Chinese version contains significant amount of information (as measured by the number of sentences)[9]. Two bilingual-speaking annotators worked to

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

[8] Section 2 described our system in terms of enriching English documents using Japanese documents. In this section, we are enriching Chinese documents using English featured articles.

[9] Article pairs with short Chinese documents are easy because the simplest solution is to translate the English article in its entirety; on the other hand, lengthy documents on both sides is a likely indicator of distributed editing.

Table II. Articles selected for manual inspection and sentence alignment. The table shows the number of sentences in the English edition (#EN) and the Chinese edition (#ZH). The "Aligned" column shows the number of sentences in English that are aligned to some sentences in Chinese and "%New" indicates the percentage of English sentences that are considered new information. The column "A(1/2/3+)" shows the percentage of English sentences that align to 1, 2, and 3 or more Chinese sentences, respectively. The column "Parallel?" indicates whether the Chinese version is created as an exact parallel translation or not, based on manual inspection of edit histories.

Article	**#EN**	**#ZH**	**Aligned**	**%New**	**A(1/2/3+)**	**P**arallel?
Acetic acid	194	169	155	20%	95/4/1%	Some
Angkor Wat	149	222	71	52%	89/10/1%	No
Australia	258	229	72	72%	86/11/3%	No
Ayumi Hamasaki	227	306	114	50%	92/7/1%	No
Battle of Cannae	221	149	100	55%	91/9/0%	No
Boeing 747	356	185	298	16%	98/2/0%	Yes
H II region	116	81	103	11%	99/1/0%	Yes
India	245	156	67	73%	87/10/3%	No
Knights Templar	156	119	39	75%	85/15/0%	No

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. More specifically, the annotator is instructed to read each sentence from the English edition of a selected article and identify the corresponding alignment to the Chinese side, if any. 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.

The amount of manual effort is similar to what a Wikipedia editor would have to do to facilitate cross-lingual document enrichment. It is a laborious process since on average the featured articles selected have 210 sentences in one English document and substantial amounts in Chinese. If the document structure of both versions are significantly different, significant mental effort is required to scan for new information. 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. In other words, despite its laboriousness, information disparity as defined here is a well-defined task.

5.2. Analysis of Information Disparity

First, we discuss how information disparity manifests itself on Wikipedia based on analyzing the manual annotation data. Table II presents statistics and observations from the annotation. Note that these article pairs are relatively rich on both sides. On average, an English featured article has 212 sentences, 46 paragraphs, and 13 sections, and the Chinese counterpart has 178 sentences, 50 paragraphs, and 16 sections. Here we define "section" by the third-level heading tag in Wikipedia, which roughly corresponds to topical subsections.

There is some qualitative differences among the articles, with varying amounts of information disparity. We found that for articles with very little information disparity (i.e. the articles with low **%New**, such as`‘‘H II region’’` and `‘‘Boeing 747’’`) the Chinese version was mainly written as a parallel translation of the original English featured article. These articles exhibit similar document structure (as evidenced by similar section headings) as well as a considerable percentage of 1-to-1 English-to-Chinese sentence alignments. For example, the column A(1/2/3+) indicates that for the `‘‘Boeing 747’’` article, 98% of English sentences align to exactly 1 Chinese sentence, 2% of English sentences align to exactly 2 Chinese sentences, and 0% of English sentences align to 3 or more Chinese sentences.

Table III. AUC Results for Task 1 (Identifying new information) and Section Accuracy Results for Task 2 (Cross-lingual sentence insertion) on manually-annotated Wikipedia articles.

Article	Task 1				Task 2		
	Maxsim	SVM	LM	Rand	Manual	Heuristic	Graph
Acetic Acid	70.8	79.6	29.1	24.3	85.7	92.8	100
Angkor Wat	81.3	86.4	69.3	49.8	66.6	83.3	66.6
Australia	92.9	93.1	79.0	74.7	50.0	50.0	66.6
Ayumi Hamasaki	72.5	72.3	58.3	50.1	90.0	70.0	100
Battle of Cannae	84.6	83.1	64.4	54.6	100	66.6	100
Boeing 747	54.1	54.1	24.5	19.2	79.3	62.0	75.8
H II Region	54.9	71.3	14.7	46.8	60.0	80.0	90.0
India	95.4	95.7	81.5	71.2	100	100	80.0
Knights Templar	89.3	93.6	84.0	79.1	66.6	33.3	66.6
AVERAGE	77.3	81.0	56.1	52.2	77.6	70.9	82.9

On the other hand, for article pairs with considerable information disparity (e.g. `Angkor Wat`), there are fewer 1-1 alignments and the document structure is very different, due to independent contributions in different language communities. Also, the annotators spent much more time in annotating these structurally-diverging article pairs, since more mental effort is required to detect new information. There are also article pairs that are between the two extremes (e.g. ``` ``Acetic Acid'' ```), which appear to be created by both periods of parallel translation effort and independent editing.

Qualitatively we found that the Wikipedia meta-data such as edit history, discussion log, and table-of-contents structure are quite indicative of the kind of information disparity existing in actual article pairs. Although we do not use this meta-data in our experiments, we imagine that they could be leveraged in interesting ways to further improve our system.

5.3. Identifying New Information

Firstly, we report our experiment of identifying sentences that contain new information. Our test set contains the nine articles that are 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:

(1) **Maxsim**: One of our proposed method that operates under the assumption that new information has low maximum cosine similarity (Section 3.1). We use the Google Translate service as the MT engine (which returns a single 1best) since it has wide-coverage.
(2) **SVM**: SVM classifier with 30% partial labels (Section 3.2). In particular, for each article, we assume there are labels for 15% of sentences with the highest Maxsim and 15% of sentences with lowest MaxSim values.
(3) **LM**: Novelty detection using Language Models (LM). One common method for novelty detection in the statistics literature [**?**] is to fit a parametric model on the data of interest; a test sample is judged novel if it has low likelihood (high perplexity) with respect to the model. Here we experimented with n-gram LMs fitted on the Chinese translations. English sentences with high perplexity (normalized by sentence length) are judged as new information.
(4) **Rand**: Random ranking of English sentences, where top ranks correspond to new information. This serves as a sanity check.

We evaluate the performance of the above methods using the *area under the precision-recall curve (AUC)* for each annotated document. Precision/recall is preferred over other measures such as ROC (Receiver operating characteristic) because of the skew in the labels. For the random method, AUC will be 50% for balanced data, $>50\%$ for articles with more new information, and $<50\%$ for articles with less new informa-

tion. We prefer to use AUC and evaluate the entire ranking of results, since this is more general than evaluating classification accuracy, whose results critically depends on classifier thresholds. Furthermore, a ranking evaluation is appropriate if we intend to use our system as an interactive assistant for a human editor. Nevertheless, we should also note that while AUC is best for summarizing a ranking of results, a system with higher AUC may not necessarily win in precision-recall for a particular setting of the classifier threshold.

Task 1 results are shown in Table III. On average, Maxsim achieves 77.3% AUC and SVM achieves 81%. Both outperform the LM baseline of 56% (this was obtained using 3-grams with Witten-Bell discount, which was the best parameter setting for LM). These relatively high values imply that current MT performance and our proposed unsupervised and partially-supervised solutions are already of sufficient quality for real-world data.

5.4. Cross-lingual Sentence Insertion

Next, we describe our experiments on the sentence insertion task. The manual alignments provide ground truth for the positioning of sentences. First, we randomly select a English sentence that has an alignment (and thus position) to the Chinese side. Then we cover up the alignment and delete the Chinese counterpart, effectively turning the English sentence into new information. The task is therefore to infer where the English sentence should go when translated into Chinese.[10] Here, we cover 50% of the alignments and measure performance in terms of *Section Accuracy*, defined as the percentage of times the new information is correctly placed in the correct section. Other evaluation metrics are possible: *Paragraph Accuracy* measures whether the new information is inserted into the relevant paragraph and *Sentence Distance* measures how many sentences are between the predicted and correct position. On average, the target side has 16 sections and 50 paragraphs, so random prediction would give 6% and 2% section and paragraph accuracies. Since our goal is to assist human editors, methods giving high section/paragraph accuracy and low sentence distance can greatly narrow down the reading one needs to do in order to enrich the target document.

We test the performance of the following three methods:

(1) **Manual**: Heuristic insertion using manual alignment references (Section 4.1). This is a oracle result of the heuristic method, assuming a perfect error-free cross-lingual similarity metric.
(2) **Heuristic**: Heuristic insertion using MT-based similarity metric (using Google Translate). (Section 4.1)
(3) **Graph**: Graph-based method using MT-based similarity metric. (Section 4.2)

The results are shown in Table III. On average, **Graph** achieves 82.9% accuracy and is more robust than the heuristic method using the same similarity information (**Heuristic**: 70.9%). In some cases, the Graph method even outperforms the heuristic using manual alignments (**Manual**), implying that global document structure is very helpful in practice. The same **Graph** system achieves *Paragraph Accuracy* of 76% and *Sentence Distance* of 11.3 [Au Yeung et al. 2011]; we imagine this performance is already sufficient for helping editors quickly identify and evaluate how a new information fits into the discourse structure of the article to be enriched.

[10]Rather than covering-up alignments, an alternative evaluation for Task 2 would be to directly annotate where a genuinely new English sentence should be placed in the Chinese version. However, this poses significant costs on the annotation process.

Table IV. Error types, example sentences, and number of False Positive (FP) and False Negative (FN) classified according to each error type.

Error Type	Example English Wikipedia sentence and article name	Corresponding sentence in Chinese version (machine translated)	FP	FN
Poor Translation	(Battle of Cannae): Ordinarily each of the two consuls would command their own portion of the army, but since the two armies were combined into one, the Roman law required them to alternate their command on a daily basis.	When will the two consuls were directing their department, but this time by two military one, so in response to the request of the Roman law, the two consuls during the day turns to command.	13	4
Lexical Mismatch	(Acetic Acid): Another 1.5 Mt are recycled each year, bringing the total world market to 6.5 Mt/a.	Annual world consumption of 6.5 million tons, the remaining 1.5 million tons were recycled.	11	0
Spurious Matching	(Australia): Separate colonies were created from parts of New South Wales: South Australia in 1836, Victoria in 1851, and Queensland in 1859.	1851 Beisesite New South Wales, Victoria Balalete discovery of gold, free settlers began to surge.	0	18
Partial Information	(Australia): After sporadic visits by fishermen from the immediate north, and European discovery by Dutch explorers in 1606, the eastern half of Australia was claimed by the British in 1770 and initially settled through penal transportation to the colony of New South Wales, founded on 26 January 1788.	Jan. 26, 1788, English navigator Arthur. Philip (Captain Arthur Phillip) led the first settlers to settle in Sydney, and raised the British flag, Australia officially became a British colony.	24	22
Contradiction	(Australia): Australia ranks 7th overall in the Center for Global Development's 2008 Commitment to Development Index	And global human development index ranking second (2009)	2	6

5.5. Error Analysis

Finally, we perform an error analysis to understand the frequent sources of mistakes. For Task 1, we manually inspected 100 English sentences which are deemed as False Positives (FP) or False Negatives (FN) according to our Maxsim method. In order to compute FP and FN, we need to set a threshold to the Maxsim values in order to reduce the evaluation to a classification problem. We chose this threshold for each article based on the amount of new information shown in Table II. Based on our observations of the data, we divided the errors into the following types:

— **Poor translation:** The translation result (from Chinese-to-English) was poor, thus the Maxsim metric was unreliable from the first stage.
— **Lexical mismatch:** The translation is semantically correct, but the words do not match the existing information. This is the fault of using a simple lexical matching similarity such as cosine and could be alleviated if one incorporates synonym or paraphrase knowledge. (This leads to False Positive, i.e. something identified as new even though there are existing information.)
— **Spurious matching:** This is the inverse of the above, where topical words common in many sentences match under cosine similarity (despite the *tf-idf* scheme), so genuinely new information may be misjudged as existing (False Negative).

— **Partial information:** The Chinese sentence contains only part of the information in the English sentence (and vice versa). In our annotation, we consider something as new information only if the sentence is entirely new, but "partially-new" sentences are prevalent in practice.
— **Contradiction:** The Chinese and English sentences may be such that one may entail the other but not vice versa (i.e. general vs. specific), or may be simply contradictory. Our annotation guideline indicates this as new information, but it may be difficult for a automatic system to discern.

Table IV shows the number of FP and FN identified for each error and example sentences. Errors due to Partial Information are the most prevalent, accounting for half of both FP and FN. Partial information has an interesting side-effect on the cosine similarity: since cosine is normalized by sentence length, sentence pairs with partial information overlap (usually of very different lengths) tend to have their cosine similarities penalized. So, the issue of whether partial information should be considered novel remains an important open question. For FP, the remaining errors are divided between Poor Translation and Lexical Mismatch. This could be fixed by better machine translation or better lexical metrics: i.e. the Lexical Mismatch example in Table IV could be solved if "Mt" and "million tons" were known as synonyms. For FN, Spurious Mismatch and Contradictions are the main sources of errors. Spurious Mismatch of named entities (e.g. "New South Wales") were especially common. Contradiction problems occur because two sentences may match in the majority of words but contain contradictory key information. Our annotation guidelines prefer to label contradictions as new information in order to alert the human editor of potential problems. Based on this error analysis, we think that the most important problems for this task going forward would be (1) rigorous definition of partial information, and (2) better cross-lingual similarity metrics to reduce to amount of Poor Translation, Lexical Mismatch, and Spurious Matching.

We also performed an error analysis for Task 2. In Table III, while **Graph** performs best overall, it appears that the individual accuracies vary by article. So one question is whether differences among **Graph**, **Manual**, and **Heuristic** could depend on translation quality, document structure, or amount of new information. As it turned out, we could not find any noticeable correlation with per-article accuracy, though it does seem that inaccurate cross-lingual similarity (due to Partial Information or Spurious Matching in particular) is an important cause of error. Furthermore, we tried a t-test on the sentence level and found that **Graph** indeed outperforms **Heuristic** by statistically significant margins ($p < 0.05$).[11] We therefore believe that looking for differences among articles may be a red herring. **Graph** looks at the entire cross-lingual similarity matrix and can be considered a generalization of **Heuristic**, which only looks at the previous similarity values: so it is conceivable that **Graph** is better in general and worse only in cases when Spurious Matching in far-away locations causes an error for **Graph**.

6. EXPERIMENTS WITH LARGE-SCALE SIMULATIONS

Experimental evaluation at a large scale is one of the main contributions of this work. While Section 5 focuses on a real dataset, here we use large simulated data in order to systematically investigate how our system performs under different conditions. In particular, we use bilingual document collections (i.e. *bitext*) commonly used in machine translation research and simulate information disparity by deleting sentences

[11] Significance testing on the article-level is not possible due to insufficient samples. The difference between sentence-level and article-level is analogous to macro-average and micro-average accuracies.

on target article. Using bitext enables us to experiment in large-scale since we avoid the laborious annotation process of Section 5.

Our goals in this section are:

(1) To evaluate whether our methods are robust when we break the system assumption to varying degrees. In particular, we focus on the assumption of "sentence as the unit of information."
(2) To understand how effective is the MT-based cross-lingual similarity (Section 2.2) on the overall system, in particular by examining variants (e.g. changing MT engine quality and incorporating MT N-best results) and other metrics (e.g. based on textual entailment and topic models).

In the following, we first explain how we prepare the data to simulate information disparity, then present a series of results and discussions.

6.1. Data Preparation

6.1.1. Data collection. We use as bitext the NICT Japanese-English Corpus of Wikipedia's Kyoto Articles, containing about 500k sentence pairs in 14k articles.[12] These are Wikipedia articles originally written in Japanese on topics related to Kyoto tourism, traditional culture, history, and religion. Each Japanese sentence is translated by hand into English. Note that the English is an exact translation of the Japanese, not the English Wikipedia version on the same topic.

Eighty percent of the data is used for training a machine translation (MT) system, as required by the cross-lingual sentence similarity computation. A total of 2517 articles (amounting to 78k original sentence pairs) is used for cross-lingual document enrichment experiments. Data statistics are shown in Table V. Note that these datasets were randomly divided along articles (not sentences), so that there may exist some domain mismatch between the topics in MT training set, evaluation set, and document enrichment set.

Table V. Statistics of various data used in the experiments. Some data were filtered following standard MT pre-processing procedure.

DATASET	#articles	#sentences	#words(JP)	#words(EN)
Machine Translation Training	11,274	285k	4.9M	5.1M
Machine Translation Tuning	147	4k	96k	100k
Machine Translation Evaluation	147	4k	101k	104k
Document Enrichment Evaluation	2,517	78k	1.8M	1.9M

6.1.2. MT System Setup. We built our own machine translation (MT) system in order to examine the effects of MT errors on the overall system. Our MT is a statistical phrase-based system, trained using the Moses toolkit [Koehn et al. 2007]. We built an Japanese-to-English system, though either translation direction is suitable in our framework. The system uses word alignments with IBM Model 4 [Brown et al. 1993], grow-diag-final-and heuristic for phrase extraction [Och and Ney 2004], MSD lexical models for reordering, trigram language models by SRILM [Stolcke 2002], and minimum error rate training [Och 2003] on the BLEU metric [Papineni et al. 2002]. This achieved 17.70 (uncased) BLEU on our MT evaluation dataset. Although the BLEU score is not high (due to the challenge of long-distance reordering in phrase-based models), the unigram precision of 53.2% on single reference seems passable for our purpose of computing cosine distance. We also artificially created lower quality MT

[12]Available at http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

system by reducing the MT training data; their performance are summarized in Table VI. Our cross-lingual similarity metrics based on N-best lists are computed from N-best lists of size up to 300.

Table VI. Performance of different MT systems on MT evaluation set.

% Train Data	BLEU	1-gram Precision
100%	17.70%	53.2%
50%	16.09%	51.7%
25%	14.21%	49.9%

6.1.3. Alternative Cross-lingual Similarity Metrics. We implemented the following cross-lingual similarity metrics as comparison to the basic MT+cosine metric:

— **MT+Entailment** (S_{entail}): This metric follows the idea of [Mehdad 2010], which uses MT and then monolingual textual entailment inference. A sentence is considered new information if it does not entail any other sentence. We use the EDITS open-source software [Kouylekov and Negri 2010] as our entailment engine. It predicts an entailment if the edit-distance operations on words or parse trees of two sentences is small. Here we use word edit distance, whose optimal edit costs are trained using the supplied genetic algorithm. The training set consists of entailment pairs generated from sentence-aligned bitext (MT Evaluation dataset of Table V), where aligned sentences represent positive entailment, and randomly-paired non-aligned sentences represent false entailment. The cross-lingual similarity is then defined as the the probability of entailment given by EDITS.

— **Polylingual Topic Model** (S_{topic}): This is our re-implementation of [Mimno et al. 2009], which fits a Bayesian model to a comparable bilingual dataset. MT is not required. The generative process is summarized as follows: For an article pair, we first draw a topic distribution θ from a Dirichlet prior, then draw the actual latent topic assignments (from 100 topics) for English ($z^e \sim$ Multinomial($z^e|\theta$)) and Japanese ($z^j \sim$ Multinomial($z^j|\theta$)). Finally, English words are generated from distributions based on z^e while Japanese words are generated from distributions based on z^j. To compute cross-lingual similarity, we infer the topic proportions per sentence and calculate the Hellinger distance $\frac{1}{2}\sum(\sqrt{P(z^j)} - \sqrt{P(z^e)})^2$, following [Blei and Lafferty 2007]. The important point is that we can obtain this model from comparable (not parallel) bitext, and that an MT engine is not used.

6.1.4. Information Disparity Setup. In this section, we focus on *enriching English* articles with Japanese articles. To simulate information disparity, we randomly perturb each article in the cross-lingual document enrichment dataset in the following way:

(1) Randomly delete d fraction of English sentences.
(2) Randomly concat c fraction of neighboring English sentences.
(3) Identify the section boundaries in the English document. Randomly shuffle the sections.

We varied $d = \{.3, .5\}$ and $c = \{.1, .2, .3\}$ in order to examine how our methods work under a variety of conditions. Large d means that there is much less information in English compared to Japanese. Large c implies fewer one-by-one correspondence between sentences, which is more realistic in multilingual document collections authored by non-corresponding parties. This tests one of the main assumptions of our system, which is that sentences are the main units of information. Section shuffling further makes this task more realistic by assuming that the general document structure may

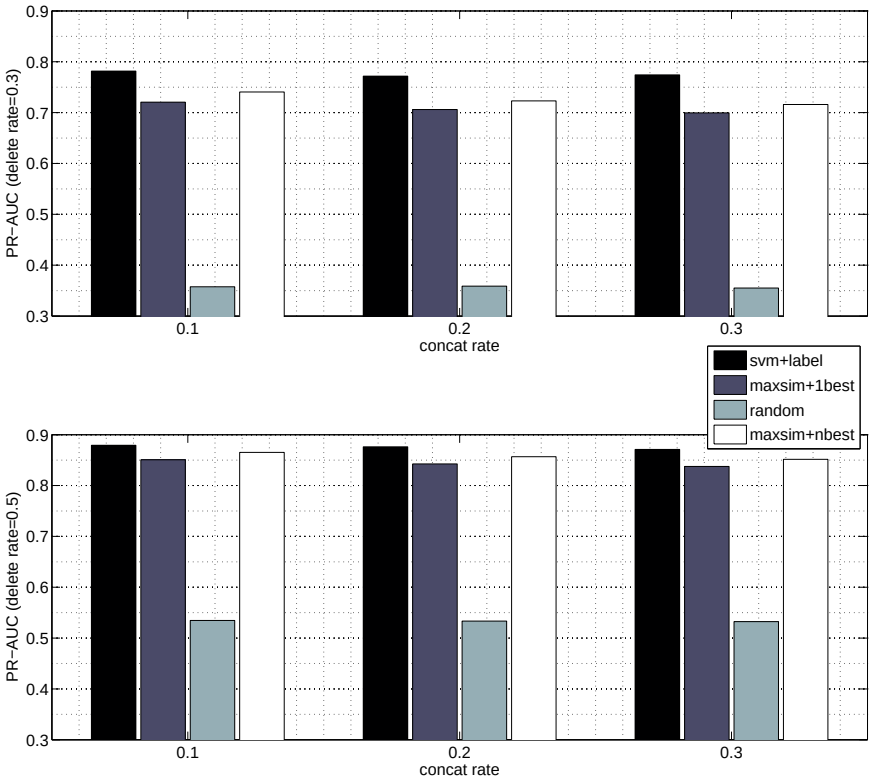


Fig. 5. New Information Detection: AUC for delete rate = {0.3, 0.5} (top/bottom) and concat rate = {0.1,0.2,0.3} (left/middle/right).

diverge among languages. We do not, however, shuffle smaller units such as paragraphs or sentences because we believe this may destroy the coherence and legibility of the articles.

For each experimental condition, we repeat the deletion/concatenation/shuffling process for 5 random trials. Our results here report the average of 5 random trials.

6.2. Identifying New Information

6.2.1. How does performance vary under different conditions?. We compare four systems under different concat and delete conditions:

— **Maxsim+1best**: proposed method using MT 1-best in similarity metric (S_{1best})
— **Maxsim+Nbest**: proposed method using MT N-best in similarity metric (S_{prob})
— **Random**: random prediction
— **SVM+label**: SVM classifier using 20% partial labels.

Figure 5 shows the AUC under six different conditions. First, observe the results for the proposed method **Maxsim+1best**: For the $d = 0.3$ condition, it achieves 72% AUC under $c = 0.1$ and degrades only slightly to 70% under the harsher $c = 0.3$; for the $d =$

0.5 condition, it achieves 85% AUC under $c = 0.1$ and degrades slightly to 84% under $c = 0.3$. So, our assumption of sentence as the unit of information seems valid for Task 1: increasing c which merges multiple sentences only degrades performance slightly, though there is indeed a noticeable correlation between c and final performance.

Next, observe that other methods exhibit similar curves for varying c. Using N-best lists (**Maxsim+Nbest**) slightly outperforms 1-best, with 71-74% AUC. The **SVM+label** results show that we can improve AUC by around 7-9% if some labels are available. As expected, the AUC results of **Random** are close to the actual amount of information deleted d. The actual AUC is not equal to d but slightly higher since concatenation reduces the number of sentences slightly.

To summarize, for Task 1, our systems achieves 70-80% AUC range when 30% of article is new, and 80-90% AUC when half of the article is new. Further, the proposed Maxsim and SVM methods are relatively robust to cases where the "sentence as unit of information" assumption does not hold, though we do notice a correlation.

6.2.2. How do different similarity metrics compare?. We now perform a more in-depth evaluation of the different definitions of cross-lingual similarity. Table VII shows the AUC of different similarity definitions, when paired with either the MaxSim or SVM method.[13]

First, note that the basic MT-1best with cosine similarity (S_{1best}) achieves 76.2% AUC (with MaxSim). N-best lists have the potential to substantially improve upon this, as evidenced by 83.0% for S_{oracle}; S_{cat} and S_{prob} is able to improve 1-2% AUC upon 1-best. It appears that S_{cat} slightly outperforms S_{prob}; this suggests that the increased vocabulary coverage by the N-best list may be a more important factor than actual probability/confidence values of translation candidates.

Second, observe that degraded MT does affect performance to some degree: For an MT trained on 50% of bitext, we observe a BLEU degradation of 1.6 leading to an AUC degradation of 76.2 - 73.0 = 3.2%. Further reducing the MT training data to 25%, we see a BLEU degradation of 3.49 leading to an AUC degradation of 5.7%.

Finally, we found that S_{entail} and S_{topic} by themselves do not give good AUC, though are quite helpful when combined (linearly-summed) with the MT-based cosine similarity S_{cat}. $S_{cat} + S_{topic} + S_{entail}$ achieves the best results, 79.2% with MaxSim and 84.4% with SVM.

Overall, our conclusion is that Task 1 is indeed sensitive to the reliability of similarity values and enhancements using N-best, better MT, or orthogonal information (such as entailment or topic models) can lead to noticeable improvements.

6.3. Cross-Lingual Sentence Insertion

6.3.1. How does performance vary under different conditions?. We compare five systems under a variety of sentence concatenation (c) and deletion (d) conditions:

— **Manual**: Heuristic insertion based on manual references (oracle).
— **Graph+nbest**: Graph method using similarity from MT N-best lists, using S_{prob}.
— **Graph+1best**: Graph method using similarity from MT 1-best result (S_{1best}).
— **Graph+1best-smalldata**: Graph method using similarity from 1-best result of MT trained on 25% data
— **Heuristic+1best**: Heuristic insertion using same similarity as **Graph+1best**

[13]The AUC numbers here are evaluated on the ''Culture'' subset of the Kyoto Wikipedia corpus (365 articles) with conditions $c = 0.3$ and $d = 0.5$. The numbers in Table VII are thus not directly comparable to Figure 5, which evaluates on the entire 2517-article set, though we expect result trends to be similar. The reason for using a smaller subset here is because of the computational cost of training pairwise entailment pairs and topic models on large datasets.

Table VII. Comparison of cross-lingual similarity for Task 1. The numbers indicate average AUC(%) ± standard deviation. The results are ranked in order of MaxSim AUC.

Cross-lingual similarity used	**Prediction method**	
	MaxSim	SVM
S_{oracle}: MT nbest-oracle, cosine	83.0 ± 1.5	85.9 ± 1.4
$S_{cat} + S_{topic} + S_{entail}$	79.2 ± 1.5	84.4 ± 1.3
$S_{cat} + S_{topic}$	78.0 ± 1.5	83.0 ± 1.4
S_{cat}: MT nbest-concat, cosine	77.6 ± 1.5	82.0 ± 1.5
$S_{cat} + S_{entail}$	77.2 ± 1.5	82.3 ± 1.5
S_{prob}: MT nbest-prob, cosine	77.0 ± 1.6	82.1 ± 1.4
S_{1best}: 1best, cosine	76.2 ± 1.6	81.0 ± 1.5
$S_{1best-smalldata}$ degraded MT w/ 50% data, cosine	73.0 ± 1.6	78.5 ± 1.5
$S_{1best-smalldata}$ degraded MT w/ 25% data, cosine	70.5 ± 1.6	76.3 ± 1.6
S_{topic}: Polylingual topic model	68.6 ± 1.6	73.9 ± 1.6
S_{entail}: MT 1-best with Entailment probability	63.5 ± 1.5	72.0 ± 1.6
Random	53.4 ± 1.4	-

Figure 6 shows Section Accuracies under six different conditions. First, observe the results of **Graph+1best**: For $d = 0.3$, accuracy degrades by 0.6% (from 93.2% to 92.6%) as we increase concat rate c from $c = 0.1$ to $c = 0.3$; for $d = 0.5$, accuracy degrades by 0.7% (from 91.3% to 90.6%) for $c = 0.1$ to $c = 0.3$. On the other hand, using the same similarity metric, **Heuristic+1best** degrades much more drastically: for $d = 0.3$, accuracy degrades by 1.9% (from 89.2% to 87.3%) as concat rate increases from $c = 0.1$ to $c = 0.3$; similarly for $d = 0.5$, accuracy degrades by 2% (from 86.4% to 84.4%). Thus a graph-based approach that incorporates soft alignments and global structure is much more robust to cases where the "sentence as unit of information" assumption is broken.

Second, note that **Manual**, which uses true alignment links as cross-lingual similarity, outperforms both **Graph+nbest** and **Graph+1best** in all six conditions. This implies that a better cross-lingual similarity has much potential to further improve an automatic system.

To summarize, our overall conclusion for Task 2 is: (a) Section accuracies around 90-95% can be achieved with all conditions (Paragraph accuracies, not shown, are around 65-70% range), and (b) using global structure such as graphs is very helpful in allowing a graceful degradation when our "sentence as unit" assumption is somewhat violated.

6.3.2. How do different similarity metrics compare?. We now observe how various similarity metrics compare, when paired with **Heuristic** and **Graph** methods. We also included **Heuristic Reverse**, which is similar to **Heuristic** but uses the successive rather than preceding alignments for finding insertion positions. Table VIII shows the systems ranked by Section Accuracies.

Similar to findings for Task 1, we see that the combination of $S_{cat} + S_{topic} + S_{entail}$ gives the best results (89.0% accuracy with Graph), outperforming individual metrics. In fact, the difference between this and S_{oracle} is quite small, suggesting that the textual entailment engine and polylingual topic models can help much in case of MT errors.

A more detailed analysis of MT error's effects can be seen by comparing S_{1best} with $S_{1best-smalldata}$. A BLEU degradation of 1.6 (50% MT bitext) leads to an accuracy degradation of 87.8 - 86.0 = 1.8%. Further reducing the MT training data to 25%, we see a BLEU degradation of 3.49 leading to an accuracy degradation of 3.4%.

In summary, similar to Task 1, we found that high c does give a noticeable degradation but the reliability of cross-lingual similarity metrics appear to be even more important.

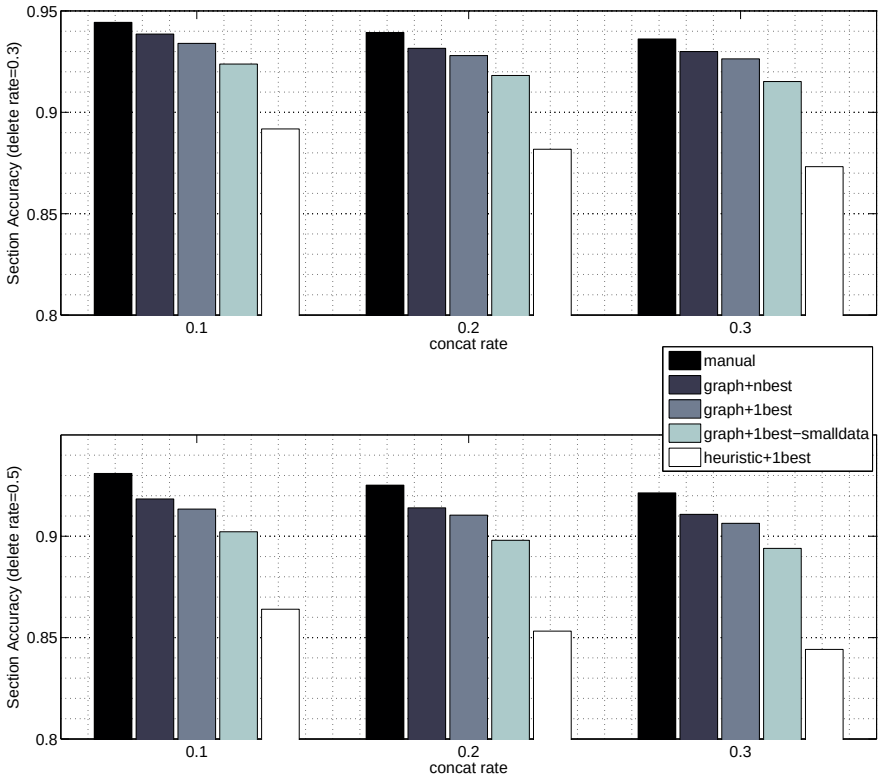


Fig. 6. Cross-lingual Insertion: Section Accuracy for delete rate = {0.3, 0.5} (top/bottom) and concat rate = {0.1,0.2,0.3} (left/middle/right).

Table VIII. Comparison of Cross-lingual similarity for Task 2. The numbers indicate Section Accuracy ± standard deviation.

Cross-lingual similarity used	**Insertion method**		
	Heuristic	Heuristic Reverse	Graph
S_{oracle}: MT nbest-oracle, cosine	80.7 ± 2.7	78.9 ± 2.8	90.1 ± 1.6
$S_{cat} + S_{topic} + S_{entail}$	81.2 ± 2.7	79.9 ± 2.8	89.0 ± 1.6
Manual	88.9 ± 1.8	88.8 ± 1.7	-
S_{prob}: MT nbest-prob, cosine	81.0 ± 2.7	80.0 ± 2.7	88.0 ± 1.8
S_{cat}: MT nbest-concat, cosine	80.9 ± 2.7	79.7 ± 2.8	87.9 ± 1.8
S_{1best}: 1best, cosine	80.8 ± 2.7	79.6 ± 2.8	87.8 ± 1.9
$S_{cat} + S_{entail}$	80.6 ± 2.7	79.6 ± 2.8	86.6 ± 2.0
$S_{1best-smalldata}$ MT w/ 50% data, cosine	80.1 ± 2.8	79.9 ± 2.7	86.0 ± 2.1
$S_{1best-smalldata}$ MT w/ 25% data, cosine	79.3 ± 2.9	79.3 ± 2.9	84.4 ± 2.2
$S_{cat} + S_{topic}$	80.6 ± 2.7	79.8 ± 2.8	82.6 ± 2.2
S_{topic}: Polylingual topic model	74.2 ± 4.0	73.3 ± 4.0	80.7 ± 3.4
S_{entail}: MT 1-best w/ Entailment probability	71.8 ± 4.1	70.5 ± 4.2	81.4 ± 4.0

6.4. Significance Tests and Final Recommendation

We have performed various experiments with different combinations of cross-lingual similarity metric, new information identification method, and insertion method. Fi-

nally, we give some final recommendations to summarize the best system we would use in practice. For Task 1, we recommend MaxSim as it is a simple yet robust method. In an interactive scenario where the user provides feedback about new information, the SVM method gives a nice improvement. For Task 2, the Graph method, which generalizes the Heuristic, gives consistently better results and is recommended. The most important factor in both tasks, however, is not the method per se but the underlying cross-lingual similarity metric. We believe that most gains could be achieved by improving the metric to be more robust to translation errors, lexical mismatch, and issues relating to partial information.

Table IX summarizes the significance results (paired t-test on articles) of the similarity metric in both tasks. We see that S_{all} which combines multiple information sources from translation N-best lists, topic models, and textual entailment outperforms all other metrics in both tasks. The differences between S_{prob} and S_{cat} are statistically not significant, while their improvements over S_{1best} is only significant for Task 1.

Table IX. Summary of significance test results for the cross-lingual similarity metric on both tasks. For each cell, {1,2} indicates that the row metric outperforms the column metric by statistically-significant margins for Tasks 1 and 2, respectively. x indicates "not statistically significant" at level $p < 0.05$.

	S_{all}	S_{prob}	S_{cat}	S_{1best}	S_{topic}	S_{entail}
$S_{all} = S_{cat} + S_{topic} + S_{entail}$	-	1,2	1,2	1,2	1,2	1,2
S_{prob}: MT nbest-prob, cosine	-	-	x,x	1,x	1,2	1,2
S_{cat}: MT nbest-concat, cosine	-	-	-	1,x	1,2	1,2
S_{1best}: 1best, cosine	-	-	-	-	1,2	1,2
S_{topic}: Polylingual topic model	-	-	-	-	-	1,2
S_{entail}: MT 1-best w/ Entailment	-	-	-	-	-	-

7. RELATED WORKS

7.1. Information Management Systems

In general, the field of managing multi-lingual collections is still relatively new. There are a few projects with similar motivations (i.e. reducing information disparity), though the problem setups are considerably different from ours.

First, along the lines of enriching semi-structured data on Wikipedia, Adar et al. [Adar et al. 2009] introduce an automated system called Ziggurat, which can be used to align and complement infoboxes across different languages. The authors build a classifier to judge whether two entries from infoboxes in different languages refer to the same thing, based on a set of features such as word similarity and out-going links. In related work, the DBpedia project [Auer et al. 2007; Auer and Lehmann 2007] aims at extracting information from infoboxes, links and categories in order to create structured data.

Another line of work focuses on cross-lingual link discovery (see, for example, the NTCIR CrossLink Evaluation Campaign[14]). Links among documents are important in reflecting the relationships between terms and entities. The goal is to discover salient links between documents regardless of the language of writing. For example, [Sorg and Cimiano 2008; Knoth et al. 2011] generalize the explicit semantic analysis method of [Gabrilovich and Markovitch 2007] to cross-lingual settings. These methods can potentially be used as plug-in replacement for our cross-lingual similarity metric.

The most related work to ours is perhaps the EU CoSyne project[15] [Monz et al. 2011]. The goal is to automatically synchronize multi-lingual Wikipedia, and in a sense, it is

[14]http://ntcir.nii.ac.jp/CrossLink/
[15]http://www.cosyne.eu/

a much more ambitious than our work of cross-lingual enrichment. [Monz et al. 2011] identifies four steps in this process: (1) pinpointing topically related information, (2) identifying new information, (3) translating, and (4) insertion in the appropriate place. Our work can be considered as tackling only step (2) and step (4), while assuming sentence as the unit of information in step (1) and assuming a human translator (not MT) will work on step (3).

Within this CoSyne project, [Mehdad et al. 2010; 2011] propose to identify new information using cross-lingual textual entailment[16]; this allows for bidirectional enrichment, since entailment prediction can be in either direction. This allows the CoSyne project to handle multi-lingual information fusion, as opposed to the one-directional enrichment we setup here. Further work by [Negri et al. 2011] discusses how one can create a dataset for cross-lingual textual entailment using crowdsourcing techniques. The idea is to ask annotators to paraphrase, simplify, or extend some sentence (to generate entailment pairs), then translate in order to obtain cross-lingual pairs. This suggests an interesting alternative to our large-scale simulation studies. Finally, [Gaspari et al. 2011] show positive responses from human editors who work with MT output. We assume many of the techniques and results presented here would be helpful in the context of the CoSyne framework too. For example, our graph-based sentence insertion method could benefit their step (4).

7.2. Component Technologies

Our system currently uses relatively straightforward methods, e.g. cosine; we believe many advanced NLP technologies could potentially be plugged-in to benefit the overall system.

Our first task is to identify sentences that contain new information when comparing two documents in different languages. A related task is to determine the similarity between two documents written in different languages. For example, [Pinto et al. 2009] propose to apply the IBM model 1 [Brown et al. 1993] to various cross-lingual NLP tasks, such as text classification, information retrieval and plagiarism detection. In particular, cross-lingual plagiarism detection [Barrp-cedeno et al. 2008] focuses on identifying similar texts in different languages on the sentence level. On the other hand, Adafre and de Rijke [Adafre and de Rijke 2006] present experiments on finding similar sentences across different languages in Wikipedia. The authors use similar methods as ours to compute cross-lingual sentence similarity. The work differs in that (1) the intended application is information retrieval and question answering, and (2) they do not evaluate MT N-best lists as they use an online MT service.

Our second task is to identify suitable positions for inserting sentences that contain new information. Related problems such as sentence ordering and alignment have been studied in natural language processing, and it is possible that our work can benefit from techniques in this area. For example, Lapata [Lapata 2003] proposes using a Markov chain to model the structure of a document. On the other hand, Barzilay and Elhadad [Barzilay and Elhadad 2003] proposes a method for sentence alignment that involves first matching larger text fragments by clustering and further refine these matches to find sentence alignments using local similarity measures. These techniques, however, usually require training to be performed on a large corpus. In contrast, our proposed model operates only on the article level and does not require any labels.

In the monolingual Wikipedia setting, [Chen et al. 2007] propose an interesting algorithm to insert new information into existing texts using data about past user edits. Sentences are represented by lexical, positional and temporal features, and the

[16]See the related SemEval task: http://www.cs.york.ac.uk/semeval-2012/task8/

weights of different features are learnt in order to calculate the scores of nodes in the document tree for sentence insertion. We do not exploit edit histories in this work, though we believe similar methods could potentially improve the cross-lingual insertion task as well.

Finally, some works focus directly on the text generation. Sauper and Barzilay [Sauper and Barzilay 2009] propose a method for generating Wikipedia articles. Their idea is to first induce an article template automatically from articles on similar topics. Relevant texts are then retrieved from the Web and a trained model is used to determine which sentences should be put under which sections. 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.

8. DISCUSSION AND CONCLUSIONS

In this paper, we propose a framework for managing information disparity in multilingual document collections, formulating the problem as *cross-lingual document enrichment*. The main challenges were to identify sentences that contain new information, and suggest positions of insertion. We showed that our unsupervised methods utilizing machine translation and graph-based methods could achieve reasonable performance. We performed two evaluations, first demonstrating a proof-of-concept feasibility of the proposed framework by evaluating against manual annotations on a real-world dataset, then systematically investigating how the system performs under various stress tests.

We summarize our conclusions as follows:

— On real-world data, reasonable performance (i.e. 77% AUC in Task 1, 82% Section Accuracy in Task 2) can be achieved with unsupervised methods. Although the results are not perfect enough for full automation of information disparity management, they already suggest it is feasible to build an interactive assistive interface for human editors.
— On large-scale simulations, we found that the system degrades gracefully when the assumption of "sentence as the unit of information" is broken. No doubt a harsh concatenation rate such as $c = 0.3$ can degrade results, but this can be remedied by building more robust algorithms. For example in Task 2 results, Heuristics degrade by 2% accuracy while Graph-based methods degrade by only 0.7% accuracy under high c.
— We find that cross-lingual similarity is the most important component of our overall system, with significant impact on final Task 1 and Task 2 performance. While MT 1-best and cosine similarity is a simple and effective solution, more advanced methods involving N-best lists, topic models, text entailment, and combinations thereof pose the most promise for improving overall performance.

It may be instructive to look at an example result: Figure 7 shows how we enriched the English Wikipedia article on 'Macau' with its Chinese version. The article is a featured (high-quality) article in Chinese but not in English, and it is a more well-known topic to the Chinese-speaking community. The figure shows three sentences identified as containing new information (A, B, C) as well as the suggested position of insertion. The first sentence (A) taken from the Chinese edition provides an alternative etymology and is a very good addition to the English document. Further, it is inserted at an appropriate location. The second sentence (B) can also be considered as new information as it elaborates on Macau's historic relationship with neighbors. In this example sentence we see that there does not seem exist a definite insertion location,

Macau

Etymology

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

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] 惟後期鋪設的碎石造工不佳，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. 7. An example of enriching the English “Macau” Wikipedia article using information from its Chinese counterpart. We show only part of the page.

though the sentence is at the correct section. Finally, the last sentence (C), while being a sentence containing new information for the English edition, is an incorrect example. The paragraph in English describes historic settlements in Macau, but the sentence is actually about a popular tourist spot in Macau nowadays. A close look at the result reveals that the translated sentence has a high similarity to the sentence just in front of the insertion position because they both contain the word ‘stone’, which is a rare word throughout the documents. In fact, since this Chinese sentence refers to a topic

(the tourist spot) that is not present in the English edition, it becomes very difficult for the algorithm to find a correct position.

We imagine such a system can be very helpful in assisting human editors to manage information disparity in multilingual collections. The system can suggest sentences that are possibly new information, and when these are placed *in context* within the target document, the editor can quickly evaluate whether this piece of information is worth translating.

As multi-lingual collections become increasingly prevalent in the future, the challenge of managing information disparity becomes more pertinent. We think much more research can contribute in this area, both within our *cross-lingual document enrichment* framework as well as other novel frameworks such as collaborative editing [Kumaran et al. 2010; Huberdeau et al. 2008] and multi-lingual synchronization [Monz et al. 2011].

REFERENCES

Fissaha Adafre Adafre and Maarten de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. 62–69.

Eytan Adar, Michael Skinner, and Daniel S. Weld. 2009. Information arbitrage across multi-lingual Wikipedia. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 94–103. DOI:http://dx.doi.org/10.1145/1498759.1498813

Ching-Man Au Yeung, Kevin Duh, and Masaaki Nagata. 2011. Providing cross-lingual editing assistance to Wikipedia editors. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II (CICLing'11)*. Springer-Verlag, Berlin, Heidelberg, 377–389. http://portal.acm.org/citation.cfm?id=1964750.1964786

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*. Springer, 722–735.

Sören Auer and Jens Lehmann. 2007. What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In *ESWC '07: Proceedings of the 4th European conference on The Semantic Web*. Springer-Verlag, Berlin, Heidelberg, 503–517. DOI:http://dx.doi.org/10.1007/978-3-540-72667-8_36

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Corinna Cortes, and Mehryar Mohri. 2009. Polynomial Semantic Indexing. In *NIPS*.

Alberto Barrp-cedeno, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On cross-lingual plagiarism analysis using a statistical model. In *Proceedings of the ECAI'08 PAN Workshop*. 9–13.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, Morristown, NJ, USA, 25–32. DOI:http://dx.doi.org/10.3115/1119355.1119359

David Blei and John Lafferty. 2007. A correlated topic model of Science. *Annals of Applied Statistics* 1, 1 (2007), 17–35.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19 (June 1993), 263–311. Issue 2. http://portal.acm.org/citation.cfm?id=972470.972474

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* (2006).

Erdong Chen, Benjamin Snyder, and Regina Barzilay. 2007. Incremental Text Structuring with Online Hierarchical Ranking. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 83–91. http://www.aclweb.org/anthology/D/D07/D07-1009

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41, 6 (1990).

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*.

Federico Gaspari, Antonio Toral, and Sudip Kumar Naskar. 2011. User-focused Task-oriented MT Evaluation for Wikis: A Case Study.. In *Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators*.

Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*. ACM, New York, NY, USA, 291–300. DOI:http://dx.doi.org/10.1145/1753326.1753370

L.P. Huberdeau, S. Paquet, and A. Desilets. 2008. The Cross-Lingual Wiki Engibe: Enabling Collaboration Across Language Barriers. In *Proc. of WikiSym*.

T. Joachims. 2006. Training Linear SVMs in linear time. In *KDD*.

Petr Knoth, Lukas Zilka., and Zdenek Zdrahal. 2011. Using Explicit Semantic Analysis for Cross-Lingual Link Discovery. In *5th International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011)*.

P. Koehn and others. 2007. Moses: open source toolkit for statistical machine translation. In *ACL*.

Milen Kouylekov and Matteo Negri. 2010. An Open-Source Package for Recognizing Textual Entailment. In *ACL Demonstrations*.

A. Kumaran, N. Datha, B. Ashok, K. Saravanan, A. Ande, A. Sharma, S. Vedantham, V. Natampally, V. Dendi, and S. Maurice. 2010. WikiBABEL: A System for Multilingual Wikipedia Content. In *Proc. of AMTA Workshop on Collaborative Translation: Technology, Crowdsourcing and the translator perspective*.

Mirella Lapata. 2003. Probabilistic text structuring: experiments with sentence ordering. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 545–552. DOI:http://dx.doi.org/10.3115/1075096.1075165

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Markos Markou and Sameer Singh. 2003. Novelty detection: A review, part 1: Statistical approaches. *Signal Processing* 83 (2003), 2481–2497.

Yashar Mehdad. 2010. Automatic cost estimation for tree edit distance using particle swarm optimization. In *ACL (short paper)*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards Cross-Lingual Textual Entailment. In *Proceedings of NAACL*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment. In *ACL*.

David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*.

Christof Monz, Vivi Nastase, Matteo Negri, Angela Fahrni, Yashar Mehdad, , and Michael Strube. 2011. CoSyne: A Framework for Multilingual Content Synchronization of Wikis. In *Proceedings of the International Symposium on Wikis and Open Collaboration*.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In *EMNLP*.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *ACL*.

Franz Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics* (2004).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*.

David Pinto, Jorge Civera, Alberto Barrón-Cedeòo, Alfons Juan, and Paolo Rosso. 2009. A statistical approach to crosslingual natural language tasks. *J. Algorithms* 64 (January 2009), 51–60. Issue 1. DOI:http://dx.doi.org/10.1016/j.jalgor.2009.02.005

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *In 37th Annual Meeting of the Association for Computational Linguistics*.

Christina Sauper and Regina Barzilay. 2009. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 208–216.

Philipp Sorg and Philipp Cimiano. 2008. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes for the CLEF 2008 Workshop*.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *ICSLP*.

X. Zhu, Z. Ghahramani, and J. Lafferty. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proceedings of International Conference on Machine Learning*.

Received February 2007; revised March 2009; accepted June 2009