Unsupervised Synthesis of Multilingual Wikipedia Articles

Chen Yuncong
The Human Language Technology Center
The Hong Kong University of Science and Technology
ee_cyxab@stu.ust.hk

Pascale Fung
The Human Language Technology Center
The Hong Kong University of Science and Technology
pascale@ee.ust.hk

Abstract

In this paper, we propose an unsupervised approach to automatically synthesize Wikipedia articles in multiple languages. Taking an existing high-quality version of any entry as content guideline, we extract keywords from it and use the translated keywords to query the monolingual web of the target language. Candidate excerpts or sentences are selected based on an iterative ranking function and eventually synthesized into a complete article that resembles the reference version closely. 16 English and Chinese articles across 5 domains are evaluated to show that our algorithm is domain-independent. Both subjective evaluations by native Chinese readers and ROUGE-L scores computed with respect to standard reference articles demonstrate that synthesized articles outperform existing Chinese versions or MT texts in both content richness and readability. In practice our method can generate prototype texts for Wikipedia that facilitate later human authoring.

1 Introduction

Wikipedia has over 260 versions in different languages, but the great disparity in their scope and quality is hindering the effective spread of knowledge. The English version is currently the dominant one with over 3 million articles while the Chinese version, for example, has only one tenth the amount. Most Chinese articles suffer from content incoherence and lack of details compared to their English counterparts. Some of these articles are human-authored translation of the English version with varying degrees of accuracy and completeness, and others are ill-arranged combinations of excerpts directly adapted from external sources. The former takes considerable human effort and the latter tends to produce fragmented and incomplete texts. The intuitive solution of machine translation is also not feasible because it hardly provides satisfactory readability.

These problems call for a *synthesis* approach. In order to present the information conveyed by an English article in Chinese, instead of literally translate it, we build a topic-template expressed by the keywords extracted from the English article. Machine-translation of these keywords helps to yield the topic-template in Chinese. Using the topic-template in Chinese, we form a pool of candidate excerpts by retrieving Chinese documents from the Internet. These online documents are usually human-authored and have optimal readability and coherence. Candidate excerpts are further split into segments as synthesis unit. For segment selection, we propose an iterative ranking function that aims to maximize textual similarity, keywords coverage, and content coherence, while penalizes information redundancy.

A feature of our approach is the use of bi-lingual resources throughout the synthesis process. We calculate similarity scores of two texts based on both English and Chinese versions of them, which forms a more precise measure than using either version alone.

For the sake of clarity, we will use English and Chinese as examples of source and target language respectively when describing the methodology. Nonetheless, our approach is not constrained to any specific language pair and supports both direction of synthesis.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 197–205,
Beijing, August 2010

2 Related Work

Much work has been done to explore the multilingualism of Wikipedia. (Adafre et al. 2006) investigated two approaches to identify similarity between articles in different languages for automatic generation of parallel corpus, including a machine-translation based approach and one using a bilingual lexicon derived from the hyperlink structure underlying Wikipedia articles. Both methods rely on pair-wise comparisons made at the sentential level, which hardly account for similarity or coherence in the paragraph scope. Besides it is not a generative algorithm and thus inapplicable to our problem where comparable sentences in Chinese are simply not available.

A generative approach was proposed by (Sauper and Barzilay, 2009) to create highly-structured Wikipedia articles (e.g. descriptions of diseases) composed of information drawn from the Internet. It uses an automatically-induced domain-specific template, and the perceptron algorithm augmented with a global integer linear programming (ILP) formulation to optimize both local fit of information into each section and global coherence across the entire article. This method works only for specific domains where articles have obviously separable sections (e.g. Causes and Symptoms) and it requires a training corpus for each domain to induce the template. Moreover, the synthesis units they use are complete excerpts rather than individual sentences as in our approach. Their choice is based on the assumption that texts on the Internet appear in complete paragraphs, with structure strictly adhere to the fixed training templates, which may be true for specific domains they test on, but fails to hold for domain-independent application. Instead, our algorithm aims to synthesize the article in the sentential level. We select sentences to fit the source content at run time, regardless to whether a pre-determined structural template exists or not. Therefore the requirement on the structures of source articles becomes very flexible, enabling our system to work for arbitrary domain. In a sense, rather than being a structure-aware approach, our algorithm performs in a content-aware manner. This also makes maintaining coherence throughout article a lot more challenging.

Works on monolingual extractive text summarization also lend insights into our problem. (Goldstein et al., 2000) used sequential sentence selection based on Maximal Marginal Relevance Multi-Document (MMR-MD) score to form summarizations for multiple documents, with the constraint of sentence count. Since our problem does not have this constraint, we employ a variant of MMR-MD and introduced new terms specific to this task. (Takamura and Okumura, 2009) formulated a text summarization task as a maximum coverage problem with knapsack constraint and proposed a variety of combinatorial mathematics-based algorithms for solving the optimization problem.

For multi-lingual summarization, (Evans, 2005) applied the concept of multi-lingual text similarity to summarization and improved readability of English summaries of Arabic text by replacing machine translated Arabic sentences with highly similar English sentences whenever possible.

3 Methodology

Figure 1 describes the high-level algorithm of our approach. The system takes as input the English Wikipedia page and outputs an article in Chinese.

First, the structured English article is extracted from the Wikipedia page. Due to the relative independence of contents in different sections in typical Wikipedia articles (e.g. childhood, early writings), a separate synthesis task is performed on each section and all synthesized sections are eventually combined in the original order to form the Chinese article.

For each section, keywords are extracted from the English text using both tf-idf and the graph-based TextRank algorithm. Named entities, time indicators, and terms with Wikipedia hyperlinks are also included. These keywords express the topics of the current section and are regarded as the content guideline. We then use Google Translate and Google Dictionary to

Input:

English version of an entry

Output:

Synthesized Chinese version

Algorithm:

1: Parse the English Wikipedia page to extract the structured texts.
2: For each section:
    2.1: Extract keywords.
    2.2: Use Chinese translation of keywords to search online Chinese texts.
    2.3: Filter retrieved Chinese texts and split them into segments.
    2.4: Synthesize the current section using candidate segments.
3: Generate the Chinese Wikipedia page by combining synthesized sections according to the original structure of English version.

Figure 1. High-level algorithm of the synthesis approach

obtain the Chinese translations of these keywords and thereby convert the content guideline into Chinese. The Chinese keywords are then combined with the translated subject term and section title to form queries that are used to retrieve online Chinese documents by Google search. The returned Chinese documents are clustered and filtered based on both their format and content. The remaining candidate excerpts are further split using the TextTiling algorithm (Hearst, 1997) into segments that constitutes the text units for synthesis. This unit size ensures both semantic completeness within each unit and flexibility of combining multiple units into coherent paragraphs. Segments are chosen according to scores computed iteratively by a variant of the MMR-MD scoring function that considers not only the relevance of an individual segment to the source section but also its impact on the provisional synthesized section as a whole.

3.1 Wikipedia Page Preprocessing

The source Wikipedia page is parsed to remove non-textual page elements (e.g. images, info-boxes and side-bars). Only texts and headings are extracted and their structures are maintained as templates for final integration of synthesized sections.

3.2 Keyword Extraction

The keyword set K for a section is the union of 6 categories of content-bearing terms.

$$K = \bigcup K_c$$

K_1: set of terms with high tf-idf score (top 5%)
K_2: set of terms with high TextRank score (top 5%)
K_3: set of named entities
K_4: set of temporal indicators (e.g. June, 1860)
K_5: set of terms with Wikipedia links
K_6: section title

For K_1, tf-idf scores are computed by:

$$tfidf_i = \sqrt{tf_i} \times \log\left(\frac{N}{df_i} + 1\right)$$

where tf_i is the term frequency of term *i* in the section and df_i is the document frequency of term *i* in a corpus consists of 2725 high-quality English Wikipedia articles [1], which well represent the language style of Wikipedia.

For K_2, we compute TextRank scores according to (Mihalcea and Tarau, 2004). It is a graph-based model where words as vertices recursively vote for the weights of their linked neighbors (e.g. words appear in the same sentence as them) using the formula:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

[1] http://evanjones.ca/software/wikipedia2text.html

Where $In(V_i)$ is the set of vertices with forward links to i, $Out(V_i)$ is the set of vertices receiving links from i, w_{ji} is the weight of edge between V_i and V_j. In the case of a word graph, we simplify this formula by assuming the graph to be undirected and unweighted. Each pair of words occurring in the same sentence share an edge between them and all word vertices have initial weights of 1.

Unlike tf-idf which considers only word-specific values and tends to give higher weights for rare words, TextRank uses global information about how a word is used in its context to induce its importance and has the advantage of highlighting keywords that are relatively common but highly relevant. In this sense, these two measures complement each other. Named entities are recognized using the named entity chunker provided by the NLTK (Natural Language ToolKit) package[2].

3.3 Keyword Translation

Keywords are then translated using Google Dictionary to form Chinese queries. Usually one English keyword has several translations and they will be used jointly when forming the search query.

Google Dictionary often fails to generate correct transliteration for rare names, so we augment it with a function of parenthesized phrase translation. We basically seeks named-entity strings from online documents that are in the format of '*CHINESE* (*ENGLISH*)' and extracts the Chinese transliteration from the pattern using regular expression combined with a Pinyin (Chinese Romanization) [3] /English pronunciation lookup table. Since Chinese words are not spaced in documents, the Pinyin/English lookup is helpful to determine the boundary of the Chinese transliteration based on the fact that most Chinese transliterations start with characters pronounced similar to the initial syllables in corresponding English names. This function is relatively simple but works surprisingly well as many rare named entities are available in this pattern on the Web.

3.4 Web Search

Keywords in Chinese alternatively form query pairs with the Wikipedia subject term. Each pair is used to retrieve a set of (16 in our experiments) Chinese documents containing both words with Google Search. If a keyword has multiple translations, they are joined by the string 'OR' in the query which is the way to specify alternatives in Google logic. If a keyword is a named entity, its English version is also used as an alternative in order to acquire documents in which the subject is referred to by its English name instead of transliterations. For the subject "Chekhov/契诃夫", a keyword with two transliterations "Taganrog/塔甘罗格/塔干罗格" and another keyword with two transliterations "father/父亲/爸爸" will result in two query pairs: "Chekhov OR 契诃夫 Taganrog OR 塔甘罗格 OR 塔干罗格" and "Chekhov OR 契诃夫 父亲 OR 爸爸".

3.5 Candidate Filtering

The retrieved excerpts are filtered first by criteria on format include text length and the percentage of white-space and non-Chinese characters. Pair-wise similarity is then computed among all the remaining excerpts and those above a certain threshold are clustered. Within a cluster only the centroid excerpt with maximum similarity with the source section will be selected. This stage typically eliminates ¾ of the documents that are either not sufficiently relevant or redundant. The similarity measure we use is a combination of both English and Chinese versions of cosine similarity and Jaccard index.

$$\mathrm{SIM}(a,b) = 0.3 \times \mathrm{COS}_{\mathrm{EN}}(a,b) + 0.3 \times \mathrm{COS}_{\mathrm{CH}}(a,b) + 0.2 \times \mathrm{JAC}_{\mathrm{EN}}(a,b) + 0.2 \times \mathrm{JAC}_{\mathrm{CH}}(a,b)$$

For Chinese excerpts, English similarity is computed by first translating them into English by Google Translate and taking tf-idf as token weights. Similar procedure works for computing Chinese similarity for English excerpts, except that Chinese texts need to be

[2] The package is available at http://www.nltk.org
[3] Pinyin information is obtained from Unicode Han Database at http://www.unicode.org/reports/tr38/

segmented[4] first and weights are based on tf only. These machine translations do not require grammatical correctness since they are essentially used as bags of words in both cosine similarity and Jaccard index. During this stage, every excerpt acquires bi-lingual versions, which is important for the extended similarity measure in the iterative ranking function.

Filtered excerpts are further split into segments using the TextTiling algorithm. After clustering the remaining segments form the candidate units for synthesis of the current section.

3.6 Iterative Scoring Function

Based on the idea that the 'goodness' of a segment should be evaluated both on its individual relevance to the source and the overall impact on the synthesized section, we summarize four factors for scoring a segment: (1) Intuitively a segment scores higher if it has higher similarity to the source section; (2) A segment makes positive contribution to synthesized section if it introduces some keywords mentioned in the source; (3) A segment tends to improve the coherence of synthesized section if it comes from the same excerpts as the other segments in synthesized section; (4) A sentence should be penalized if its content is redundant with the synthesized section.

Integrating the four factors above, we propose that for source text *r*, the score of the *i*th candidate segment s_i in the *n*th iteration is formulated as:

$$Q_{r,n}(s_i) = w_s \times S_r(s_i) + w_k \times K_{r,n}(s_i) + w_c \times C_n(s_i) - w_R \times R_n(s_i)$$

This formula is composed of 4 terms corresponding to the 'goodness' factors: $S_r(s_i)$ for similarity, $K_{r,n}(s_i)$ for keyword coverage, $C_n(s_i)$ for coherence, and $R_n(s_i)$ for redundancy. The corresponding weights are tuned in a large number of experiments as to achieve optimal performance. This function is a variant of the original MMR-MD score tailored for our application.

$S_r(s_i)$ is a comprehensive similarity measure of segment s_i to the reference text *r*.

$$S_r(s_i) = w_1 \times SIM(s_i, \mathrm{r}) + w_2 \times SIM(s_i, \mathrm{p}) + w_3 \times SIM(e_i, \mathrm{r}) + w_4 \times SIM(e_i, \mathrm{p})$$

where p is the parent section of r and e_i is the parent excerpt of s_i. Similarities between parent excerpts are also examined because sometimes two segments, especially short segments, despite their textual similarity actually come from very different contexts and exhibit different focuses. In this case, the latter three terms will suppress the score between these two segments which would otherwise be erroneously high and therefore produce a more precise measure of similarity.

$\mathrm{K_{r,n}(s_i)}$ measures the contribution of $\mathrm{s_i}$ in terms of uncovered keywords.

$$\mathrm{K_{r,n}(s_i)} = \sum_{\substack{\mathrm{k} \in \mathrm{U_{r,n}} \\ \mathrm{k} \neq \mathrm{subject}}} \mathrm{idf(k)}$$

$$\mathrm{U_{r,n}} = \mathrm{K_r} - \bigcup_{\mathrm{s_j} \in \mathrm{D_n}} \mathrm{K_j}$$

where $\mathrm{D_n}$ is the winner set in the nth iteration. $\mathrm{K_r}$ is the set of keywords in the reference text and $\mathrm{K_j}$ is the set of keywords in the selected segment $\mathrm{s_j}$. $\mathrm{U_{r,n}}$ represents the set of keywords in the reference that are not yet been covered by the provisional synthesized text in the nth iteration. $\mathrm{K_{r,n}(s_i)}$ quantifies the keyword contribution as the sum of idf values of uncovered keywords. The subject term is excluded because it as a keyword does not reflect any topic bias and is therefore not a good indicator for coverage.

$C_n(s_i)$ is a term that reflects the coherence and readability in the synthesized text.

$$C_n(s_i) = \left|\{\mathrm{s_j} | \mathrm{e_j} = \mathrm{e_i}, \mathrm{s_j} \in D_n\}\right|$$

[4] The segmentation tool using forward maximum matching is obtained at http://technology.chtsai.org/mmseg

where e_i is the parent excerpt of s_i and e_j is the parent excerpt of s_j. Segments from the same excerpts tend to be less redundant and more coherent. Therefore candidates that share the same parent excerpts as segments in winner set are more favorable and rewarded by this term. This is a major difference from the original MMR-MD function in which sentences from different documents are favored. This is because their formula is targeted for automatic summarization where more emphasis is put on diversity rather than coherence.

$R_n(s_i)$ measures the redundancy of the synthesized text if s_i is included. It is quantified as the maximum similarity of s_i with all selected segments.

$$R_n(s_i) = \max_{s_j \in D_n} S(s_i, s_j)$$

3.7 Segment Selection Algorithm

Figure 2 describes the segment selection algorithm. Starting with a candidate set and an empty winner set, we iteratively rank the candidates by Q and in each iteration the top-ranked segment is examined. There are two circumstances a segment would be selected for the winner set:

(1) if the segment scores sufficiently high
(2) the segment does not score high enough for an unconditional selection, but as long as it introduces uncovered keywords, its contribution to the overall content quality may still overweigh the compromised similarity

In the second circumstance however, since we are only interested in the uncovered keywords, it may not be necessary for the entire segment to be included in the synthesized text. Instead, we only include the sentences in this segment that contain those keywords. Therefore we propose two conditions:

- $\boldsymbol{C_{sel\text{-}segment}}$**:** condition for selecting a segment
$$Q_{r,n}(s_{top}) > 0.8 * \mathrm{Q_{max}}$$

- $\boldsymbol{C_{sel\text{-}sentence}}$**:** condition for selecting sentences
$Q_{r,n}(\mathrm{s_{stop}}) > 0.6 * \mathrm{Q_{max}}$ **and** $K_{r,n}(\mathrm{s_{stop}}) > 0$ **and** $S_r(s_{top}) > 0.3 * \mathrm{S_{max}}$

Thresholds in both conditions are not static but dependent on the highest score of all candidates in order to accommodate diversity in score range for different texts. Finally if no more candidates are able to meet the lowered score threshold, even if they might carry new keywords, we assume they are not suitable for synthesis and return the current winner set. This break condition is formulated as $\boldsymbol{C_{break}}$**:**

- $\boldsymbol{C_{break}}$**:** condition to finish selection
$$Q_{r,n}(s_{top}) < 0.6 * \mathrm{Q_{max}}$$

Input:
- $\boldsymbol{S_n}$: candidate set in iteration *n*
- $\boldsymbol{r}$: the reference text

Define:
- $\boldsymbol{n}$: iteration index
- $\boldsymbol{D_n}$: winner set in iteration *n*
- $\boldsymbol{C_{sel\text{-}segment}}$**:** $Q_{r,n}(s_{top}) > 0.8 * \mathrm{Q_{max}}$
- $\boldsymbol{C_{sel\text{-}sentence}}$**:** $Q_{r,n}(s_{top}) > 0.6 * \mathrm{Q_{max}}$ **and** $K_{r,n}(s_{top}) > 0$ **and** $S_r(s_{top}) > 0.3 * \mathrm{S_{max}}$
- $\boldsymbol{C_{break}}$**:** $Q_{r,n}(s_{top}) < 0.6 * \mathrm{Q_{max}}$

Algorithm:

$\boldsymbol{D_n} \leftarrow \emptyset$, $\boldsymbol{n} \leftarrow 0$
while $\boldsymbol{S_n} \neq \emptyset$:
  $\boldsymbol{s_{top}} \leftarrow \arg\max_{\mathrm{s_i \in S_n}} \boldsymbol{Q_{r,n}(s_i)}$
  if $\mathrm{C_{break}}$**:**
    return $\boldsymbol{D_n}$
  else if $\mathrm{C_{sel\text{-}segment}}$**:**
    $\boldsymbol{D_n} \leftarrow \boldsymbol{D_n} + \boldsymbol{s_{top}}$
  else if $\mathrm{C_{sel\text{-}sentence}}$**:**
    $\boldsymbol{D_n} \leftarrow \boldsymbol{D_n}$ + sentences in $\boldsymbol{s_{top}}$ with the uncovered keywords
    $\boldsymbol{S_n} \leftarrow \boldsymbol{S_n} - \boldsymbol{s_{top}}$
    $\boldsymbol{n} \leftarrow \boldsymbol{n} + \mathbf{1}$

Output:
Synthesized text for the reference $\boldsymbol{r}$

Figure 2. Segment selection algorithm

4 Evaluation

4.1 Experiment Setup

We evaluate our system on 16 Wikipedia subjects across 5 different domains as listed in Table 1.

Category	Subjects
Person	Anton Chekhov Abu Nuwas Joseph Haydn Li Bai
Organization	HKUST IMF WTO
Events	Woodstock Festival Invasion of Normandy Decembrist Revolt
Science	El Nino Gamma Ray Stingray
Culture	Ceramic Art Spiderman Terrorism

Table 1. Subjects used for evaluation

The subjects are selected from "the List of Articles Every Wikipedia Should Have" [5] published by Wikimedia. These subjects are especially appropriate for our evaluation because we can (1) use a subset of such articles that have high quality in both English and Chinese as standard reference for evaluation; (2) safely assume Chinese information about these subjects is widely available on the Internet; (3) take subjects currently without satisfactory versions in Chinese as our challenge.

Human Evaluation

We presented the synthesized articles of these subjects to 5 native Chinese readers who compare synthesized articles with MT results and existing Chinese versions on Wikipedia which range from translated stubs to human-authored segments. We asked the reviewers to score them on a 5-point scale in terms of four quality indicators: structural similarity to the English version, keyword coverage, fluency, and conciseness.

Automatic Evaluation

In addition to human evaluation, we also compare synthesized articles to several high-quality Chinese Wikipedia articles using ROUGE-L (C.Y. Lin, 2004). We assume these Chinese versions are the goals for our synthesis system and greater resemblance with these standard references indicates better synthesis. ROUGE-L measures the longest common subsequence (LCS) similarity between two documents, rather than simply word overlap so it to some degree reflects fluency.

4.2 Result Analysis

Human Evaluation

Human evaluator feedbacks for articles in different categories are shown in Table 2. Machine-translated versions are judged to have the highest score for structural similarity, but erroneous grammar and word choices make their readability so poor even within sentences and therefore of no practical use.

Generally, articles synthesized by our system outperform most existing Chinese versions in terms of both structural and content similarity. Many existing Chinese versions completely ignore important sections that appear in English versions, while our system tries to offer information with as much fidelity to the English version as possible and is usually able to produce information for every section. Synthesized articles however, tend to be less fluent and more redundant than human-authored versions.

Performance varies in different domains. Synthesis works better for subjects in *Person* category, because the biographical structure provides a specific and fairly unrelated content in each section, making the synthesis less redundancy-prone. On the other hand, there is arbitrariness when organizing articles in *Event* and *Culture* category. This makes it difficult to find online text organized in the same way as the English Wikipedia version, therefore introducing a greater challenge in sentence selection for each section. Articles in the *Science* category usually include rare terminologies, and formatted texts like diagrams and formula, which impede correct translation and successful extraction of keywords.

[5] http://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have/Version_1.2

Cat.	Structural Similarity			Coverage			Fluency			Conciseness		
	Synt.	Orig.	MT	**Synt.**	Orig.	MT	**Synt.**	Orig.	MT	**Synt.**	Orig.	MT
Psn.	**2.85**	1.49	5	**2.94**	1.84	4.51	**2.71**	4.58	0.83	**1.74**	4.47	n/a
Org.	**1.96**	1.22	5	**2.51**	2.10	4.46	**2.10**	4.42	1.06	**0.99**	4.53	n/a
Evt.	**1.37**	1.13	5	**2.56**	1.94	4.40	**2.45**	4.46	0.81	**0.80**	4.40	n/a
Sci.	**2.43**	1.30	5	**2.68**	2.14	4.42	**2.53**	4.51	1.02	**1.05**	4.50	n/a
Cul.	**1.39**	1.35	5	**2.2**	2.21	4.54	**2.32**	4.54	0.94	**1.34**	4.59	n/a
Avg.	**2.02**	1.30	5	**2.58**	2.05	4.47	**2.42**	4.50	0.93	**1.22**	4.50	n/a

Table 2. Result of human evaluation against English source articles (out of 5 points; Synt: synthesized articles; Orig: the existing human-authored Chinese Wikipedia versions; MT: Chinese versions generated by Google Translate)

Automatic Evaluation

Using ROUGE-L to measure the quality of both synthesized and MT articles against human-authored standard references, we find synthesized articles generally score higher than MT versions. The results are shown in Table 3.

Category	Recall		Precision		F-score	
	Synt.	MT	Synt.	MT	Synt.	MT
Psn.	**0.48**	0.30	**0.20**	0.16	**0.28**	0.22
Org.	**0.40**	0.29	**0.16**	0.13	**0.23**	0.18
Evt.	**0.36**	0.26	0.13	**0.15**	**0.19**	0.19
Sci.	**0.31**	0.22	**0.14**	0.11	**0.19**	0.15
Cul.	**0.37**	0.27	**0.13**	0.12	**0.24**	0.17
Avg.	**0.38**	0.27	**0.15**	0.13	**0.23**	0.18

Table 3. Results of automatic evaluation against gold Chinese reference articles (Synt: synthesized articles; MT: Chinese versions generated by Google Translate)

The synthesized articles, extracted from high quality human-authored monolingual texts, are generally better in precision than the MT articles because there is less erroneous word choice or grammatical mistakes. Most synthesized articles also have higher recall than MT versions because usually a substantial portion of the high-quality Chinese excerpts, after being retrieved by search engine, will be judged by our system as good candidate texts and included into the synthesized article. This naturally increases the resemblance of synthesized articles to standard references, and thus the F-scores. Note that since our method is unsupervised, the inclusion of the standard Chinese articles underscores the precision and recall of our method.

5 Conclusion

In this paper, we proposed an unsupervised approach of synthesizing Wikipedia articles in multiple languages based on an existing high-quality version of any entry. By extracting keywords from the source article and retrieving relevant texts from the monolingual Web in a target language, we generate new articles using an iterative scoring function.

Synthesis results for several subjects across various domains confirmed that our method is able to produce satisfactory articles with high resemblance to the source English article. For many of the testing subjects that are in 'stub' status, our synthesized articles can act as either replacement or supplement to existing Chinese versions. For other relatively well-written ones, our system can help provide content prototypes for missing sections and missing topics, bootstrapping later human editing.

A weakness of our system is the insufficient control over coherence and fluency in paragraph synthesis *within* each section, new methods are being developed to determine the proper order of chosen segments and optimize the readability.

We are working to extend our work to a system that supports conversion between major languages such as German, French and Spanish. The employment of mostly statistical methods in our approach facilitates the extension. We have also released a downloadable desktop application and a web application based on this system to assist Wikipedia users.

Reference

Adafre, Sisay F. and Maarten de Rijke, "Finding Similar Sentences across Multiple Languages in Wikipedia", *Proceedings of the EACL Workshop on New Text*, Trento, Italy, 2006

Bird, Steven, E. Klein, and E. Loper, *Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 2009

Evans, David K., "Identifying Similarity in Text: Multi-Lingual Analysis for Summarization", PhD thesis, Columbia University, 2005.

Goldstein, Jade, Vibhu Mittal, Jaime Carbonell and Mark Kantrowitz, "Multi-document summarization by sentence extraction", *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 40-48, 2000

Hearst, Marti A., "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", *Computational Linguistics*, Volume 23, Issue 1, pp. 33-64, 1997

Lin, Chin-Yew, "ROUGE: A Package for Automatic Evaluation of Summaries", *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, Barcelona, Spain.

Mihalcea, Rada and Paul Tarau, "TextRank: Bringing order into texts", *Proceedings of EMNLP*, pages 404–411 Barcelona, Spain, 2004

Sauper, Christina and Regina Barzilay, "Automatically Generating Wikipedia Articles: a Structure-Aware Approach", *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 208–216, Suntec, Singapore, 2-7 August 2009.

Takamura, Hiroya and Manabu Okumura, "Text Summarization Model based on Maximum Coverage Problem and its Variant", *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 781-789, 2009