Cross-Lingual Information to the Rescue in Keyword Extraction

[1]**Chung-Chi Huang** [2]**Maxine Eskenazi** [3]**Jaime Carbonell** [4]**Lun-Wei Ku** [5]**Ping-Che Yang**
[1,2,3]Language Technologies Institute, CMU, United States
[4]Institute of Information Science, Academia Sinica, Taiwan
[5]Institute for Information Industry, Taipei, Taiwan
{[1]u901571,[4]lunwei.jennifer.ku}@gmail.com
{[2]max+,[3]jgc}@cs.cmu.edu [5]maciaclark@iii.org.tw

Abstract

We introduce a method that extracts keywords in a language with the help of the other. In our approach, we bridge and fuse conventionally irrelevant word statistics in languages. The method involves estimating preferences for keywords w.r.t. domain topics and generating cross-lingual bridges for word statistics integration. At run-time, we transform parallel articles into word graphs, build cross-lingual edges, and exploit PageRank with word keyness information for keyword extraction. We present the system, *BiKEA*, that applies the method to keyword analysis. Experiments show that keyword extraction benefits from PageRank, globally learned keyword preferences, and cross-lingual word statistics interaction which respects language diversity.

1 Introduction

Recently, an increasing number of Web services target extracting keywords in articles for content understanding, event tracking, or opinion mining. Existing keyword extraction algorithm (KEA) typically looks at articles monolingually and calculate word significance in certain language. However, the calculation in another language may tell the story differently since languages differ in grammar, phrase structure, and word usage, thus word statistics on keyword analysis.

Consider the English article in Figure 1. Based on the English content alone, monolingual KEA may not derive the best keyword set. A better set might be obtained by referring to the article and its counterpart in another language (e.g., Chinese). Different word statistics in articles of different languages may help, due to language divergence such as phrasal structure (i.e., word order) and word usage and repetition (resulting from word translation or word sense) and so on. For example, bilingual phrases "social reintegration" and "重返社會" in Figure 1 have inverse word orders ("social" translates into "社會" and "reintegration" into "重返"), both "prosthesis" and "artificial limbs" translate into "義肢", and "physical" can be associated with "物理" and "身體" in "physical therapist" and "physical rehabilitation" respectively. Intuitively, using cross-lingual statistics (implicitly leveraging language divergence) can help look at articles from different perspectives and extract keywords more accurately.

We present a system, *BiKEA*, that learns to identify keywords in a language with the help of the other. The cross-language information is expected to reinforce language similarities and value language dissimilarities, and better understand articles in terms of keywords. An example keyword analysis of an English article is shown in Figure 1. *BiKEA* has aligned the parallel articles at word level and determined the scores of topical keyword preferences for words. *BiKEA* learns these topic-related scores during training by analyzing a collection of articles. We will describe the *BiKEA* training process in more detail in Section 3.

At run-time, *BiKEA* transforms an article in a language (e.g., English) into PageRank word graph where vertices are words in the article and edges between vertices indicate the words' co-occurrences. To hear another side of the story, *BiKEA* also constructs graph from its counterpart in another language (e.g., Chinese). These two independent graphs are then bridged over nodes

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 1–6,
Baltimore, Maryland USA, June 23-24, 2014.

The English Article:
I've been in Afghanistan for 21 years. I work for the Red Cross and I'm a physical therapist. My job is to make arms and legs -- well it's not completely true. We do more than that. We provide the patients, the Afghan disabled, first with the physical rehabilitation then with the social reintegration. It's a very logical plan, but it was not always like this. For many years, we were just providing them with artificial limbs. It took quite many years for the program to become what it is now. …

Its Chinese Counterpart:
我在阿富汗已經 21 年。 我為紅十字會工作， 我是一名物理治療師。 我的工作是製作胳膊和腿 -- 恩，這不完全是事實。 我們做的還不止這些。 我們提供給患者， 阿富汗的殘疾人， 首先是身體康復, 然後重返社會。 這是一個非常合理的計劃， 但它並不是總是這樣。 多年來，我們只是給他們提供義肢。 花了很多年的程序 才讓這計劃成為現在的模樣。…

Word Alignment Information:
physical (物理), therapist (治療師), social (社會), reintegration (重返), physical (身體), rehabilitation (康復), prosthesis (義肢), …

Scores of Topical Keyword Preferences for Words:
(English) prosthesis: 0.32; artificial leg: 0.21; physical therapist: 0.15; rehabilitation: 0.08; …
(Chinese) 義肢: 0.41; 物理治療師: 0.15; 康復:0.10; 阿富汗: 0.08, …

English Keywords from Bilingual Perspectives:
prosthesis, artificial, leg, rehabilitation, orthopedic, …

Figure 1. An example *BiKEA* keyword analysis for an article.

that are bilingually equivalent or aligned. The bridging is to take language divergence into account and to allow for language-wise interaction over word statistics. *BiKEA*, then in bilingual context, iterates with learned word keyness scores to find keywords. In our prototype, *BiKEA* returns keyword candidates of the article for keyword evaluation (see Figure 1); alternatively, the keywords returned by *BiKEA* can be used as candidates for social tagging the article or used as input to an article recommendation system.

2 Related Work

Keyword extraction has been an area of active research and applied to NLP tasks such as document categorization (Manning and Schutze, 2000), indexing (Li et al., 2004), and text mining on social networking services ((Li et al., 2010); (Zhao et al., 2011); (Wu et al., 2010)).

The body of KEA focuses on learning word statistics in document collection. Approaches such as tfidf and entropy, using local document and/or across-document information, pose strong baselines. On the other hand, Mihalcea and Tarau (2004) apply PageRank, connecting words locally, to extract essential words. In our work, we leverage globally learned keyword preferences in PageRank to identify keywords.

Recent work has been done on incorporating semantics into PageRank. For example, Liu et al. (2010) construct PageRank synonym graph to accommodate words with similar meaning. And Huang and Ku (2013) weigh PageRank edges based on nodes' degrees of reference. In contrast, we bridge PageRank graphs of parallel articles to facilitate statistics re-distribution or interaction between the involved languages.

In studies more closely related to our work, Liu et al. (2010) and Zhao et al. (2011) present PageRank algorithms leveraging article topic information for keyword identification. The main differences from our current work are that the article topics we exploit are specified by humans not by automated systems, and that our PageRank graphs are built and connected bilingually.

In contrast to the previous research in keyword extraction, we present a system that automatically learns topical keyword preferences and constructs and inter-connects PageRank graphs in bilingual context, expected to yield better and more accurate keyword lists for articles. To the best of our knowledge, we are the first to exploit cross-lingual information and take advantage of language divergence in keyword extraction.

3 The BiKEA System

Submitting natural language articles to keyword extraction systems may not work very well. Keyword extractors typically look at articles from monolingual points of view. Unfortunately, word statistics derived based on a language may

be biased due to the language's grammar, phrase structure, word usage and repetition and so on. To identify keyword lists from natural language articles, a promising approach is to automatically bridge the original monolingual framework with bilingual parallel information expected to respect language similarities and diversities at the same time.

3.1 Problem Statement

We focus on the first step of the article recommendation process: identifying a set of words likely to be essential to a given article. These keyword candidates are then returned as the output of the system. The returned keyword list can be examined by human users directly, or passed on to article recommendation systems for article retrieval (in terms of the extracted keywords). Thus, it is crucial that keywords be present in the candidate list and that the list not be too large to overwhelm users or the subsequent (typically computationally expensive) article recommendation systems. Therefore, our goal is to return reasonable-sized set of keyword candidates that, at the same time, must contain essential terms in the article. We now formally state the problem that we are addressing.

Problem Statement: We are given a bilingual parallel article collection of various topics from social media (e.g., TED), an article ART^e in language *e*, and its counterpart ART^c in language *c*. Our goal is to determine a set of words that are likely to contain important words of ART^e. For this, we bridge language-specific statistics of ART^e and ART^c via bilingual information (e.g., word alignments) and consider word keyness w.r.t. ART^e's topic such that cross-lingual diversities are valued in extracting keywords in *e*.

In the rest of this section, we describe our solution to this problem. First, we define strategies for estimating keyword preferences for words under different article topics (Section 3.2). These strategies rely on a set of article-topic pairs collected from the Web (Section 4.1), and are monolingual, language-dependent estimations. Finally, we show how *BiKEA* generates keyword lists for articles leveraging PageRank algorithm with word keyness and cross-lingual information (Section 3.3).

3.2 Topical Keyword Preferences

We attempt to estimate keyword preferences with respect to a wide range of article topics. Basically, the estimation is to calculate word significance in a domain topic. Our learning process is shown in Figure 2.

(1) Generate article-word pairs in training data
(2) Generate topic-word pairs in training data
(3) Estimate keyword preferences for words w.r.t. article topic based on various strategies
(4) Output word-and-keyword-preference-score pairs for various strategies

Figure 2. Outline of the process used to train *BiKEA*.

In the first two stages of the learning process, we generate two sets of article and word information. The input to these stages is a set of articles and their domain topics. The output is a set of pairs of article ID and word in the article, e.g., (ART^e=1, w^e="prosthesis") in language *e* or (ART^c=1, w^c="義肢") in language *c*, and a set of pairs of article topic and word in the article, e.g., (tp^e="disability", w^e="prosthesis") in *e* and (tp^e="disability", w^c="義肢") in *c*. Note that the topic information is shared between the involved languages, and that we confine the calculation of such word statistics in their specific language to respect language diversities and the language-specific word statistics will later interact in PageRank at run-time (See Section 3.3).

The third stage estimates keyword preferences for words across articles and domain topics using aforementioned (*ART*,*w*) and (*tp*,*w*) sets. In our paper, two popular estimation strategies in Information Retrieval are explored. They are as follows.

tfidf. $tfidf(w)=freq(ART,w)/appr(ART',w)$ where term frequency in an article is divided by its appearance in the article collection to distinguish important words from common words.

ent. $\text{entropy}(w)= -\sum_{tp'} Pr(tp'|w)\times\log(Pr(tp'|w))$ where a word's uncertainty in topics is used to estimate its associations with domain topics.

These strategies take global information (i.e., article collection) into account, and will be used as keyword preference models, bilingually intertwined, in PageRank at run-time which locally connects words (i.e., within articles).

3.3 Run-Time Keyword Extraction

Once language-specific keyword preference scores for words are automatically learned, they are stored for run-time reference. *BiKEA* then uses the procedure in Figure 3 to fuse the originally language-independent word statistics

to determine keyword list for a given article. In this procedure a machine translation technique (i.e., IBM word aligner) is exploited to glue statistics in the involved languages and make bilingually motivated random-walk algorithm (i.e., PageRank) possible.

procedure PredictKW(ART^e,ART^c,*KeyPrefs*,*WA*,α,*N*)
//Construct language-specific word graph for PageRank
(1) $\mathbf{EW}^e$=constructPRwordGraph(ART^e)
(2) $\mathbf{EW}^c$=constructPRwordGraph(ART^c)
//Construct inter-language bridges
(3) $\mathbf{EW}=\alpha\times\mathbf{EW}^e+(1-\alpha)\times\mathbf{EW}^c$
for each word alignment (w_i^c, w_j^e) in *WA*
if IsContWord(w_i^c) and IsContWord(w_j^e)
(4a) $\mathbf{EW}[i,j]+=1\times BiWeight^{cont}$
else
(4b) $\mathbf{EW}[i,j]+=1\times BiWeight^{noncont}$
(5) normalize each row of **EW** to sum to 1
//Iterate for PageRank
(6) set $\mathbf{KP}_{1\times v}$ to
[*KeyPrefs*(w_1), *KeyPrefs*(w_2), …,*KeyPrefs*(w_v)]
(7) initialize $\mathbf{KN}_{1\times v}$ to [1/*v*,1/ *v*, …,1/*v*]
repeat
(8a) **KN**'=λ× **KN**× **EW**+(1-λ)× **KP**
(8b) normalize **KN**' to sum to 1
(8c) update **KN** with **KN**' after the check of **KN** and **KN**' until *maxIter* or avgDifference(**KN**,**KN**')≤ *smallDiff*
(9) *rankedKeywords*=Sort words in decreasing order of **KN**
return the *N rankedKeywords* in *e* with highest

Figure 3. Extracting keywords at run-time.

Once language-specific keyword preference scores for words are automatically learned, they are stored for run-time reference. *BiKEA* then uses the procedure in Figure 3 to fuse the originally language-independent word statistics to determine keyword list for a given article. In this procedure a machine translation technique (i.e., IBM word aligner) is exploited to glue statistics in the involved languages and make bilingually motivated random-walk algorithm (i.e., PageRank) possible.

In Steps (1) and (2) we construct PageRank word graphs for the article ART^e in language *e* and its counterpart ART^c in language *c*. They are built individually to respect language properties (such as subject-verb-object or subject-object-verb structure). Figure 4 shows the algorithm. In this algorithm, **EW** stores normalized edge weights for word w_i and w_j (Step (2)). And **EW** is a *v* by *v* matrix where *v* is the vocabulary size of ART^e and ART^c. Note that the graph is directed (from words to words that follow) and edge weights are words' co-occurrences within window size *WS*. Additionally we incorporate edge weight multiplier *m*>1 to propagate more PageRank scores to content words, with the intuition that content words are more likely to be keywords (Step (2)).

procedure constructPRwordGraph(*ART*)
(1) $\mathbf{EW}_{v\times v}=0_{v\times v}$
for each sentence *st* in *ART*
for each word w_i in *st*
for each word w_j in *st* where *i*<*j* and *j*-*i*≤ *WS*
if not IsContWord(w_i) and IsContWord(w_j)
(2a) $\mathbf{EW}[i,j]+=1\times m$
elif not IsContWord(w_i) and not IsContWord(w_j)
(2b) $\mathbf{EW}[i,j]+=1\times (1/m)$
elif IsContWord(w_i) and not IsContWord(w_j)
(2c) $\mathbf{EW}[i,j]+=1\times (1/m)$
elif IsContWord(w_i) and IsContWord(w_j)
(2d) $\mathbf{EW}[i,j]+=1\times m$
return **EW**

Figure 4. Constructing PageRank word graph.

Step (3) in Figure 3 linearly combines word graphs $\mathbf{EW}^e$ and $\mathbf{EW}^c$ using α. We use α to balance language properties or statistics, and *BiKEA* backs off to monolingual KEA if α is one.

In Step (4) of Figure 3 for each word alignment (w_i^c, w_j^e), we construct a link between the word nodes with the weight *BiWeight*. The inter-language link is to reinforce language similarities and respect language divergence while the weight aims to elevate the cross-language statistics interaction. Word alignments are derived using IBM models 1-5 (Och and Ney, 2003). The inter-language link is directed from w_i^c to w_j^e, basically from language *c* to *e* based on the directional word-aligning entry (w_i^c, w_j^e). The bridging is expected to help keyword extraction in language *e* with the statistics in language *c*. Although alternative approach can be used for bridging, our approach is intuitive, and most importantly in compliance with the directional spirit of PageRank.

Step (6) sets **KP** of keyword preference model using topical preference scores learned from Section 3.2, while Step (7) initializes **KN** of PageRank scores or, in our case, word keyness scores. Then we distribute keyness scores until the number of iteration or the average score differences of two consecutive iterations reach their respective limits. In each iteration, a word's keyness score is the linear combination of its keyword preference score and the sum of the propagation of its inbound words' previous PageRank scores. For the word w_j^e in ART^e, any edge (w_i^e,w_j^e) in ART^e, and any edge (w_k^c,w_j^e) in *WA*, its new PageRank score is computed as below.

$$\mathbf{KN}'[1,j] = \lambda \times \left(\begin{array}{c} \alpha \times \sum_{i \in v} \mathbf{KN}[1,i] \times \mathbf{EW}^{e}[i,j] + \\ (1-\alpha) \times \sum_{k \in v} \mathbf{KN}[1,k] \times \mathbf{EW}[k,j] \end{array} \right) + (1-\lambda) \times \mathbf{KP}[1,j]$$

Once the iterative process stops, we rank words according to their final keyness scores and return top N ranked words in language e as keyword candidates of the given article ART^{e}. An example keyword analysis for an English article on our working prototype is shown in Figure 1. Note that language similarities and dissimilarities lead to different word statistics in articles of difference languages, and combining such word statistics helps to generate more promising keyword lists.

4 Experiments

BiKEA was designed to identify words of importance in an article that are likely to cover the keywords of the article. As such, *BiKEA* will be trained and evaluated over articles. Furthermore, since the goal of *BiKEA* is to determine a good (representative) set of keywords with the help of cross-lingual information, we evaluate *BiKEA* on bilingual parallel articles. In this section, we first present the data sets for training *BiKEA* (Section 4.1). Then, Section 4.2 reports the experimental results under different system settings.

4.1 Data Sets

We collected approximately 1,500 English transcripts (3.8M word tokens and 63K word types) along with their Chinese counterparts (3.4M and 73K) from TED (www.ted.com) for our experiments. The GENIA tagger (Tsuruoka and Tsujii, 2005) was used to lemmatize and part-of-speech tag the English transcripts while the CKIP segmenter (Ma and Chen, 2003) segment the Chinese.

30 parallel articles were randomly chosen and manually annotated for keywords on the English side to examine the effectiveness of *BiKEA* in English keyword extraction with the help of Chinese.

4.2 Experimental Results

Table 1 summarizes the performance of the baseline *tfidf* and our best systems on the test set. The evaluation metrics are nDCG (Jarvelin and Kekalainen, 2002), precision, and mean reciprocal rank.

(a) @*N*=5	nDCG	P	MRR
tfidf	.509	.213	.469
PR+tfidf	.676	.400	.621
BiKEA+tfidf	.703	.406	.655

(b) @*N*=7	nDCG	P	MRR
tfidf	.517	.180	.475
PR+tfidf	.688	.323	.626
BiKEA+tfidf	.720	.338	.660

(c) @*N*=10	nDCG	P	MRR
tfidf	.527	.133	.479
PR+tfidf	.686	.273	.626
BiKEA+tfidf	.717	.304	.663

Table 1. System performance at (a) *N*=5 (b) *N*=7 (c) *N*=10.

As we can see, monolingual PageRank (i.e., *PR*) and bilingual PageRank (*BiKEA*), using global information tfidf, outperform *tfidf*. They relatively boost nDCG by 32% and P by 87%. The MRR scores also indicate their superiority: their top-two candidates are often keywords vs. the 2nd place candidates from *tfidf*. Encouragingly, *BiKEA+tfidf* achieves better performance than the strong monolingual *PR+tfidf* across *N*'s. Specifically, it further improves nDCG relatively by 4.6% and MRR relatively by 5.4%.

Overall, the topical keyword preferences, and the inter-language bridging and the bilingual score propagation in PageRank are simple yet effective. And respecting language statistics and properties helps keyword extraction.

5 Summary

We have introduced a method for extracting keywords in bilingual context. The method involves estimating keyword preferences, word-aligning parallel articles, and bridging language-specific word statistics using PageRank. Evaluation has shown that the method can identify more keywords and rank them higher in the candidate list than monolingual KEAs. As for future work, we would like to explore the possibility of incorporating the articles' reader feedback into keyword extraction. We would also like to examine the proposed methodology in a multi-lingual setting.

Acknowledgement

This study is conducted under the “Online and Offline integrated Smart Commerce Platform (1/4)” of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

Scott A. Golder and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *Information Science*, 32(2): 198-208.

Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. In *Proceedings of the WWW*, pages 211-220.

Chung-chi Huang and Lun-wei Ku. 2013. Interest analysis using semantic PageRank and social interaction content. In *Proceedings of the ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction*, pages 929-936.

Kalervo Jarvelin and Jaana Kekalainen. 2002. Cumulated gain-based evaluation of IR technologies. *ACM Transactions on Information Systems*, 20(4): 422-446.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 48-54.

Quanzhi Li, Yi-Fang Wu, Razvan Bot, and Xin Chen. 2004. Incorporating document keyphrases in search results. In *Proceedings of the Americas Conference on Information Systems*.

Zhenhui Li, Ging Zhou, Yun-Fang Juan, and Jiawei Han. 2010. Keyword extraction for social snippets. In *Proceedings of the WWW*, pages 1143-1144.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Proceedings of the ACL Workshop on Multi-Source Multilingual Information Extraction and Summarization*, pages 17-24.

Zhengyang Liu, Jianyi Liu, Wenbin Yao, Cong Wang. 2010. Keyword extraction using PageRank on synonym networks. In *Proceedings of the ICEEE*, pages 1-4.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the EMNLP*, pages 366-376.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the ACL Workshop on Chinese Language Processing*.

Chris D. Manning and Hinrich Schutze. 2000. *Foundations of statistical natural language processing*. MIT Press.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing orders into texts. In *Proceedings of the EMNLP*, pages 404-411.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-51.

Yoshimasa Tsuruoka and Jun’ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the EMNLP*, pages 467-474.

Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4): 303-336.

Wei Wu, Bin Zhang, and Mari Ostendorf. 2010. Automatic generation of personalized annotation tags for Twitter users. In *Proceedings of the NAACL*, pages 689-692.

Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. 2011. Topical keyword extraction from Twitter. In *Proceedings of the ACL*, pages 379-388.