Bilingual Keyword Extraction and its Educational Application

Chung-Chi Huang  
LTI, CMU  
5000 Forbes Ave.  
Pittsburgh, PA, USA  
u901571@gmail.com

Mei-Hua Chen  
FLL  
Tunghai University  
Taichung, Taiwan  
chen.meihua@gmail.com

Ping-Che Yang  
Institute for Information Industry  
Taipei, Taiwan  
maciaclark@iii.org.tw

Abstract

We introduce a method that extracts keywords in a language with the help of the other. The method involves estimating preferences for topical keywords and fusing language-specific word statistics. At run-time, we transform parallel articles into word graphs, build cross-lingual edges for word statistics integration, and exploit PageRank with word keyness information for keyword extraction. We apply our method to keyword analysis and language learning. Evaluation shows that keyword extraction benefits from cross-language information and language learners benefit from our keywords in reading comprehension test.

1 Introduction

Keyword extraction algorithms (KEA) have been developed to extract keywords for content understanding, event tracking, or opinion mining. However, most of them calculate article-level word keyness in a single language. The articles’ counterparts in another language may have different keyword candidates in mind since languages differ in grammar, phrase structure, and word usage, all of which play a role in word keyness statistics, thus keyword analysis.

Consider the English article in Figure 1. Monolingual KEA, based solely on the English content, may not identify the best keyword set. A better set might be obtained by consulting the article in more than a language (e.g., the Chinese counterparts) in that language divergence such as phrasal structure (i.e., word order), and word usage and word repetition (resulting from word translation or word sense) lead to different views on keywords across languages. Example English-Chinese divergence in Figure 1 includes the word order in the phrase social reintegration and 重返社會 (social translated to 社會 and reintegration inversely to 重返), many-to-one mapping/translation e.g. both prosthesis and artificial limbs translated to 義肢, and one-to-many mapping e.g. physical respectively translated to 物理 and 身體 in context physical therapist and physical rehabilitation. We hypothesize that, with the differences in languages, language-specific word statistics might be fused to contribute to keyword analysis.

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 respect 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 topical keyword preference scores for words. BiKEA learns these topic-related scores during training by analyzing a collection of articles.

At run-time, BiKEA transforms an article in a language into PageRank word graph. To hear another side of the story, BiKEA also constructs word graph from its counterpart in another language. These two graphs are then bridged over bilingually equivalent nodes. The bridging is to take language divergence into account and
to allow for language-wise interaction over word statistics. At last, BiKEA iterates in bilingual context with word keyness scores to find keywords.

2 Related Work

Keyword extraction has been actively applied to many NLP tasks: 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 (Liu et al. (2009) and Gebre et al. (2013)). On the other hand, Mihalcea and Tarau (2004) apply PageRank, connecting words locally, to extract essential words. In our work, we integrate globally learned keyword preferences into PageRank to identify keywords.

Recent work has been 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 word graphs from parallel articles to facilitate re-distribution or interaction of the word statistics of 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 automated systems, and that our PageRank graphs are built and connected bilingually.

In contrast to the previous research on topic modeling (e.g., Zhao and Xing (2007)) and keyword extraction, we present a keyword extraction algorithm that learns topical keyword preferences and bilingually inter-connects PageRank graphs. The bilinguality is to help predict better keywords taking into account the perspectives of the languages involved including the language similarities and dissimilarities. We also use our keywords for educational purpose like reading comprehension.

3 BiKEA

3.1 Problem Statement

We focus on identifying keywords of a given article in a language with the help of the other. Keyword candidates are returned as the output of the system. The returned keyword list can be examined by humans (e.g., for keyword evaluation or language learning), or passed on to article recommendation systems for article retrieval. Therefore, our goal is to return a reasonable-sized set of keyword candidates that contain the given article’s essential terms. We now formally describe 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 in language e, and its counterpart ART in language c. Our goal is to determine a set of words that are likely to contain important words of ART. For
this, we take into account word keyness w.r.t. ART’s topic and bridge language-specific statistics of ART and ARTc via bilingual information (e.g., word alignments) such that cross-lingual diversities are valued in extracting keywords in e.

3.2 Topical Keyword Preferences

We attempt to estimate language-wise 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 has following four stages.

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., (ID, ART=prosthesis) in language e or (ID, ART=prosthesis) in language c, and a set of pairs of topic and word in the article, e.g., (tp=disability, w=prosthesis) in e and (tp=disability, w=prosthesis) in c. Note that the topic information is shared across languages, and that, to respect language diversities, words’ topical significance is calculated within their specific language and the original language-independent word statistics will later be fused and interact at run-time.

The third stage estimates keyword preferences for words across articles and domain topics using aforementioned (ART, w) and (tp, w) sets. In our paper, simple yet effective tfidf estimation is used: 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.

tfidf takes global information (i.e., article collection) into account, and will be used as keyword preference model 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 learned, they are stored for run-time reference. BiKEA then uses the procedure in Figure 2 to fuse word statistics across languages to determine keyword list for a given article. In this procedure 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, ARTc.KeyPrefs, WA, A, N)
  //Construct language-specific word graph for PageRank
  (1) EW=constructPRwordGraph(ART)
  (2) EW=constructPRwordGraph(ARTc)
  //Construct inter-language bridges
  (3) EW=α×EW+(1-α)×EW for each word alignment (w1, w2) in WA
      if IsContWord(w1) and IsContWord(w2)
        EW[i,j]=1×BiWeightext
      else
        EW[i,j]=1×BiWeightinct
  (4a) EW[i,j]%=1×BiWeightext
  (4b) EW[i,j]+=1×BiWeightinct
  (5) normalize each row of EW to sum to 1
  //Iterate for PageRank
  (6) set KP_i, to
      [KeyPrefs(w1), KeyPrefs(w2), ..., KeyPrefs(wv)]
  (7) initialize KN_i, to [1/v, 1/v, ..., 1/v]
      repeat
        (8a) KN_i=λ×KN_i+EW(1-λ)×KP
        (8b) normalize KN_i to sum to 1
        (8c) update KN with KN_i after the check of KN and KN_i
        until maxIter or avgDifference(KN, KN_i) ≤ smallDiff
  (9) rankedKeywords=Sort words in decreasing order of KN
      return the N rankedKeywords in e with highest scores
```

Figure 2. Extracting keywords at run-time.

```
procedure constructPRwordGraph(ART)
  (1) EW=0,vxv for each sentence st in ART
      for each word w_i in st
        if not IsContWord(w_i) and IsContWord(w_i)
          if not IsContWord(w_i)
            EW[i,j]=1×m
          else
            EW[i,j]=1×(1/m)
            if IsContWord(w_i)
              EW[i,j]=1×(1/m)
            else
              EW[i,j]=1×(1/m)
        return EW
```

Figure 3. Constructing PageRank word graph.

In Steps (1) and (2) of Figure 2 we construct PageRank word graphs for the article ART in language e and its counterpart ARTc in language c. They are built independently using the procedure in Figure 3 to respect language properties (such as subject-verb-object or subject-object-verb structure). In the algorithm of Figure 3, 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 and ARTc. 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.
Then, Step (3) in Figure 2 linearly combines word graphs $EW^e$ and $EW^c$ using $\alpha$. We use $\alpha$ to balance language properties/statistics, and BiKEA backs off to monolingual KEA if $\alpha$ is one.

In Step (4) for each word alignment $(w^e_i, w^c_j)$, we construct a link between the word nodes with the weight $BiWeight$. The inter-language link is expected to reinforce language similarities and respect language divergence while the weight is to facilitate cross-language statistics interaction.

Word alignments $WA$ are derived using IBM models 1-5 (Och and Ney, 2003). Based on the directional word-aligned entry $(w^e_i, w^c_j)$, the inter-language link is directed from $w^e_i$ to $w^c_j$, i.e., from language $c$ to $e$. The fusion or bridging of PageRank graphs across languages 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 keyword preference model $KP$ using topical preference scores 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 $KN$ converges. 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 $w^e_i$ in $ART^e$, any edge $(w^e_i, w^c_j)$ in $ART^c$, and any edge $(w^c_i, w^e_j)$ in $WA$, its new PageRank score is computed as:

$$
KN^e[1,j] = \lambda \times \left( \alpha \times \sum_{i \in e} KN^e[1,i] \times EW^e[i,j] + \right)
+ \left(1 - \alpha \right) \times KP[1,j]
$$

Once the iterative process stops, we rank words according to their final keyness scores and return $N$ top-ranked words in language $e$ as keyword candidates of the given article $ART^e$.

4 Experiments

4.1 Data Sets

We collected 3.8M-word English transcripts along with their Chinese counterparts from TED for our experiments. GENIA tagger (Tsuruoka and Tsujii, 2005) was used to lemmatize and part-of-speech tag the English transcripts while CKIP (Ma and Chen, 2003) was used to segment the Chinese.

Fifty parallel articles (approximately 2,500 words per article) were randomly chosen and manually annotated with English keywords for keyword analysis.

4.2 Evaluation on Keywords

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

As we can see, monolingual PageRank ($PR$) and bilingual PageRank (BiKEA), using global information tfidf, outperform tfidf. They relatively boost nDCG by 21% and P by 55%. MRR’s also indicate their superiority: their top-two candidates are often keywords vs. the 2nd, ranked from tfidf. Encouragingly, BiKEA+tfidf achieves better performance than the strong monolingual $PR+tfidf$; further improving nDCG relatively by 7.4% and MRR relatively by 9.4%.

Overall, topical keyword preferences and inter-language bridging in PageRank which values language properties/statistics, help keyword extraction.

| @N=5 | P | MRR | nDCG |
|------|---|-----|-----|
| tfidf | .256 | .547 | .587 |
| $PR+tfidf$ | .396 | .663 | .712 |
| BiKEA+tfidf | .412 | .725 | .765 |

| @N=7 | P | MRR | nDCG |
|------|---|-----|-----|
| tfidf | .211 | .550 | .587 |
| $PR+tfidf$ | .337 | .669 | .720 |
| BiKEA+tfidf | .348 | .728 | .770 |

| @N=10 | P | MRR | nDCG |
|------|---|-----|-----|
| tfidf | .162 | .555 | .594 |
| $PR+tfidf$ | .282 | .669 | .719 |
| BiKEA+tfidf | .302 | .730 | .760 |

Table 1. System performance across $N$’s.

4.3 Application to Language Learning

The role of highlighting keywords in reading comprehension has been attracting interest in the field of language learning and educational psychology (Nist and Hogrebe, 1987; Peterson 1991; Silvers and Kreiner, 1997). In this paper, we further examine keywords in the context of computer assisted language learning. Specifically, we applied our automatic BiKEA to keyword highlighting in reading comprehension and intended to see how much language learners can benefit from BiKEA keywords in reading comprehension test.
In our case study, we asked an English professor to set a multiple-choice reading comprehension test based on one English TED transcript (See Figure 4) and recruited 26 second-year college students learning English as a second language. Their proficiency in English was estimated to be of pre-intermediate level.

These students were randomly and evenly divided into experimental (reading the English transcript with BiKEA keywords) and control group (reading without). Promisingly, our keywords helped the students: students in the experimental group achieved better averaged test score (.82) than those in the control group (.74). Relatively, the improvement was 10%. Moreover, post-study survey indicated that 90% of the participants found our keywords helpful for their article reading and key concept grasping. We are analyzing the influence of the highlighted BiKEA keywords on the high-performing students as well as the low-performing students in the test.

5 Summary

We have introduced a method for extracting keywords in bilingual context. The method involves automatically estimating topical keyword preferences and bridging language-specific PageRank word statistics. Evaluation shows that the method can yield better keywords than strong monolingual KEA. And a case study indicates that language learners benefit from our keywords in reading comprehension test. Admittedly, using our keywords for educational purposes needs further experiments.

Acknowledgement

This study is conducted under the "Online and Offline integrated Smart Commerce Platform (2/4)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

References

B. G. Gebre, M. Zampieri, P. Wittenburg, and T. Heskens. 2013. Improving native language identification with tf-idf weighting. In Proceedings of
the NAACL Workshop on Innovative Use of NLP for Building Educational Applications, pages 216-223.

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 Americas Conference on Information Systems.

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.

F. Liu, D. Pennell, F. Liu, and Y. Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In Proceedings of the NAACL, pages 620-628.

Zhengyang Liu, Jianyi Liu, Wenbin Yao, and Cong Wang. 2010. Keyword extraction using PageRank on synonym networks. In Proceedings of the IJCEE, 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.

S. L. Nist and M. C. Hogrebe. 1987. The role of underlining and annotating in remembering textual information. Reading Research and Instruction, 27(1): 12-25.

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

Divya Padmanabhan, Prasanna Desikan, Jaideep Srivastava, and Kashif Riaz. 2005. WICER: a weighted inter-cluster edge ranking for clustered graphs. In Proceedings of the IEEE/WIC/ACM WI, pages 522-528.

S. E. Peterson. 1991. The cognitive functions of underlining as a study technique. Reading Research and Instruction, 31(2): 49-56.

V. L. Silvers and D. S. Kreiner. 1997. The effects of pre-existing inappropriate highlighting on reading comprehension. Reading Research and Instruction, 36(3): 217-223.

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.

B. Zhao and E. P. Xing. 2007. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In Proceedings of the NIPS.

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