Engkoo: Mining the Web for Language Learning

Matthew R. Scott, Xiaohua Liu, Ming Zhou, Microsoft Engkoo Team
Microsoft Research Asia
No. 5, Dan Ling Street, Haidian District, Beijing, 100080, China
{mrscott, xiaoliu, mingzhou, engkoo}@microsoft.com

Abstract

This paper presents *Engkoo* [1], a system for exploring and learning language. It is built primarily by mining translation knowledge from billions of web pages - using the Internet to catch language in motion. Currently *Engkoo* is built for Chinese users who are learning English; however the technology itself is language independent and can be extended in the future. At a system level, *Engkoo* is an application platform that supports a multitude of NLP technologies such as cross language retrieval, alignment, sentence classification, and statistical machine translation. The data set that supports this system is primarily built from mining a massive set of bilingual terms and sentences from across the web. Specifically, web pages that contain both Chinese and English are discovered and analyzed for parallelism, extracted and formulated into clear term definitions and sample sentences. This approach allows us to build perhaps the world's largest lexicon linking both Chinese and English together - at the same time covering the most up-to-date terms as captured by the net.

1 Introduction

Learning and using a foreign language is a significant challenge for most people. Existing tools, though helpful, have several limitations. Firstly, they often depend on static contents compiled by experts, and therefore cannot cover fresh words or new usages of existing words. Secondly, their search functions are often limited, making it hard for users to effectively find information they are interested in. Lastly, existing tools tend to focus exclusively on dictionary, machine translation or language learning, losing out on synergy that can reduce inefficiencies in the user experience.

This paper presents *Engkoo*, a system for exploring and learning language. Different from existing tools, it discovers fresh and authentic translation knowledge from billions of web pages - using the Internet to catch language in motion, and offering novel search functions that allow users efficient access to massive knowledge resources. Additionally, the system unifies the scenarios of dictionary, machine translation, and language learning into a seamless and more productive user experience. *Engkoo* derives its data from a process that continuously culls bilingual term/sentence pairs from the web, filters noise and conducts a series of NLP processes including POS tagging, dependency parsing and classification. Meanwhile, statistical knowledge such as collocations is extracted. Next, the mined bilingual pairs, together with the extracted linguistic knowledge, are indexed. Finally, it exposes a set of web services through which users can: 1) look up the definition of a word/phrase; 2) retrieve example sentences using keywords, POS tags or collocations; and 3) get the translation of a word/phrase/sentence.

While *Engkoo* is currently built for Chinese users who are learning English, the technology itself is language independent and can be extended to support other language pairs in the future.

We have deployed *Engkoo* online to Chinese internet users and gathered log data that suggests its

[1] http://www.engkoo.com.

Proceedings of the ACL-HLT 2011 System Demonstrations, pages 44–49,
Portland, Oregon, USA, 21 June 2011.

utility. From the logs we can see on average 62.0% of daily users are return users and 71.0% are active users (make at least 1 query); active users make 8 queries per day on average. The service receives more than one million page views per day.

This paper is organized as follows. In the next section, we briefly introduce related work. In Section 3, we describe our system. Finally, Section 4 concludes and presents future work.

2 Related Work

Online Dictionary Lookup Services. Online dictionary lookup services can be divided into two categories. The first mainly relies on the dictionaries edited by experts, e.g., Oxford dictionaries [2] and Longman contemporary English dictionary [3]. Examples of these kinds of services include iCiba [4] and Lingoes [5]. The second depends mainly on mined bilingual term/sentence pairs, e.g., Youdao [6]. In contrast to those services, our system has a higher recall and fresher results, unique search functions (e.g., fuzzy POS-based search, classifier filtering), and an integrated language learning experience (e.g., translation with interactive word alignment, and photorealistic lip-synced video tutors).

Bilingual Corpus Mining and Postprocessing. Shi et al. (2006) uses document object model (DOM) tree mapping to extract bilingual sentence pairs from aligned bilingual web pages. Jiang et al. (2009b) exploits collective patterns to extract bilingual term/sentence pairs from one web page. Liu et al. (2010) proposes training a SVM-based classifier with multiple linguistic features to evaluate the quality of mined corpora. Some methods are proposed to detect/correct errors in English (Liu et al., 2010; Sun et al., 2007). Following this line of work, *Engkoo* implements its mining pipeline with a focus on robustness and speed, and is designed to work on a very large volume of web pages.

3 System Description

In this section, we first present the architecture followed by a discussion of the basic components; we then demonstrate the main scenarios.

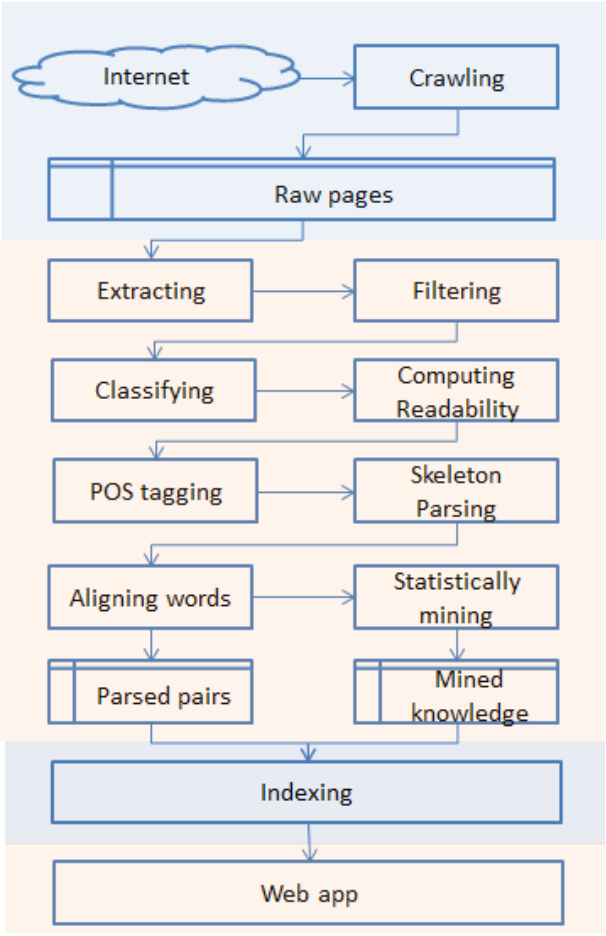

Figure 1: System architecture of *Engkoo*.

3.1 System Overview

Figure 1 presents the architecture of *Engkoo*. It can be seen that the components of *Engkoo* are organized into four layers. The first layer consists of the crawler and the raw web page storage. The crawler periodically downloads two kinds of web pages, which are put into the storage. The first kind of web pages are parallel web pages (describe the same contents but with different languages, often from bilingual sites, e.g., government sites), and the second are those containing bilingual contents. A list of seed URLs are maintained and updated after each round of the mining process.

The second layer consists of the extractor, the filter, the classifiers and the readability evaluator, which are applied sequentially. The extractor scans the raw web page storage and identifies bilingual

[2] http://oxforddictionaries.com
[3] http://www.ldoceonline.com/
[4] http://dict.en.iciba.com/
[5] http://www.lingoes.cn/
[6] http://dict.youdao.com

web page pairs using URL patterns. For example, two web pages are parallel if their URLs are in the form of "$\cdots$/zh/$\cdots$" and "$\cdots$/en/$\cdots$", respectively. Following the method of Shi et al. (2006) the extractor then extracts bilingual term/sentence pairs from parallel web pages. Meanwhile, it identifies web pages with bilingual contents, and mines bilingual term/sentence pairs from them using the method proposed by Jiang et al. (2009b). The filter removes repeated pairs, and uses the method introduced by Liu et al. (2010) to single out low quality pairs, which are further processed by a noisy-channel based sub-model that attempts to correct common spelling and grammar errors. If the quality is still unacceptable after correction, they will be dropped. The classifiers, i.e., oral/non-oral, technical/non-technical, title/non-title classifiers, are applied to each term/sentence pair. The readability evaluator assigns a score to each term/sentence pair according to Formula 1 [7].

$$206.835 - 1.015 \times \frac{\#words}{\#sentences} - 84.6 \times \frac{\#syllables}{\#words} \quad (1)$$

Two points are worth noting here. Firstly, a list of top sites from which a good number of high quality pairs are obtained, is figured out; these are used as seeds by the crawler. Secondly, bilingual term/sentence pairs extracted from traditional dictionaries are fed into this layer as well, but with the quality checking process ignored.

The third layer consists of a series of NLP components, which conduct POS tagging, dependency parsing, and word alignment, respectively. It also includes components that learn translation information and collocations from the parsed term/sentence pairs. Based on the learned statistical information, two phrase-based statistical machine translation (SMT) systems are trained, which can then translate sentences from one language to the other and vice versa. Finally, the mined bilingual term/sentence pairs, together with their parsed information, are stored and indexed with a multi-level indexing engine, a core component of this layer. The indexer is called multi-level since it uses not only keywords but also POS tags and dependency triples (e.g., "Tobj$\backsim$watch$\backsim$TV", which means "TV" is the object of "watch") as lookup entries.

The fourth layer consists of a set of services that expose the mined term/sentence pairs and the linguistic knowledge based on the built index. On top of these services, we construct a web application, supporting a wide range of functions, such as searching bilingual terms/sentences, translation and so on.

3.2 Main Components

Now we present the basic components of *Engkoo*, namely: 1) the crawler, 2) the extractor, 3) the filter, 4) the classifiers, 5) the SMT systems, and 6) the indexer.

Crawler. The crawler scans the Internet to get parallel and bilingual web pages. It employs a set of heuristic rules related to URLs and contents to filter unwanted pages. It uses a list of potential URLs to guide its crawling. That is, it uses these URLs as seeds, and then conducts a deep-first crawling with a maximum allowable depth of 5. While crawling, it maintains a cache of the URLs of the pages it has recently downloaded. It processes a URL if and only if it is not in the cache. In this way, the crawler tries to avoid repeatedly downloading the same web page. By now, about 2 billion pages have been scanned and about 0.1 parallel/bilingual pages have been downloaded.

Extractor. A bilingual term/sentence extractor is implemented following Shi et al. (2006) and Jiang et al. (2009b). It works in two modes, mining from parallel web pages and from bilingual web pages. Parallel web pages are identified recursively in the following way. Given a pair of parallel web pages, the URLs in two pages are extracted respectively, and are further aligned according to their positions in DOM trees, so that more parallel pages can be obtained. The method proposed by Jiang et al. (2007) is implemented as well to mine the definition of a given term using search engines. By now, we have obtained about 1,050 million bilingual term pairs and 100 million bilingual sentence pairs.

Filter. The filter takes three steps to drop low quality pairs. Firstly, it checks each pair if it contains any malicious word, say, a noisy symbol. Secondly, it adopts the method of Liu et al. (2010) to estimate the quality of mined pairs. Finally, following the work related to English as a second language (ESL) errors detection/correction (Liu et al., 2010; Sun et

[7] http://www.editcentral.com/gwt1/EditCentral.html

al., 2007), it implements a text normalization component based on the noisy-channel model to correct common spelling and grammar errors. That is, given a sentence $s^{'}$ possibly with noise, find the sentence $s^{*} = argmax_{s}\, p(s)p(s^{'}|s)$, where $p(s)$ and $p(s^{'}|s)$ are called the language model and the translation model, respectively. In *Engkoo*, the language model is a 5-gram language model trained on news articles using SRILM (Stolcke, 2002), while the translation model is based on a manually compiled translation table. We have got about 20 million bilingual term pairs and 15 million bilingual sentence pairs after filtering noise.

Classifiers. All classifiers adopt SVM as models, and bag of words, bi-grams as well as sentence length as features. For each classifier, about 10,000 sentence pairs are manually annotated for training/development/testing. Experimental results show that on average these classifiers can achieve an accuracy of more than 90.0%.

SMT Systems. Our SMT systems are phrase-based, trained on the web mined bilingual sentence pairs using the GIZA++ (Och and Ney, 2000) alignment package, with a collaborative decoder similar to Li et al. (2009). The Chinese-to-English/English-to-Chinese SMT system achieves a case-insensitive BLUE score of 29.6% / 47.1% on the NIST 2008 evaluation data set.

Indexer. At the heart of the indexer is the inverted lists, each of which contains an entry pointing to an ordered list of the related term/sentence pairs. Compared with its alternatives, the indexer has two unique features: 1) it contains various kinds of entries, including common keywords, POS taggers, dependency triples, collocations, readability scores and class labels; and 2) the term/sentence pairs related to the entry are ranked according to their qualities computed by the filter.

3.3 Using the System

Definition Lookup. Looking up a word or phrase on *Engkoo* is a core scenario. The traditional dictionary interface is extended with a blending of web-mined and ranked term definitions, sample sentences, synonyms, collocations, and phonetically similar terms. The result page user experience includes an intuitive comparable tabs interface described in Jiang et al. (2009a) that effectively exposes differences between similar terms. The search experience is augmented with a fuzzy auto completion experience, which besides traditional prefix matching is also robust against errors and allows for alternative inputs. All of these contain inline micro translations to help users narrow in on their intended search. Errors are resolved by a blend of edit-distance and phonetic search algorithms tuned for Chinese user behavior patterns identified by user study. Alternative input accepted includes Pinyin (Romanization of Chinese characters) which returns transliteration, as well as multiple wild card operators.

Take for example the query "tweet," illustrated in Figure 2(a). The definitions for the term derived from traditional dictionary sources are included in the main definition area and refer to the noise of a small bird. Augmenting the definition area are "Web translations," which include the contemporary use of the word standing for micro-blogging. Web-mined bilingual sample sentences are also presented and ranked by popularity metrics; this demonstrates the modern usage of the term.

Search of Example Sentences. *Engkoo* exposes a novel search and interactive exploration interface for the ever-growing web-mined bilingual sample sentences in its database. Emphasis is placed on sample sentences in *Engkoo* because of their crucial role in language learning. *Engkoo* offers new methods for the self-exploration of language based on the applied linguistic theories of "learning as discovery" and Data-Driven Learning (DDL) introduced by Johns (1991). One can search for sentences as they would in traditional search engines or concordancers. Extensions include allowing for mixed input of English and Chinese, and POS wild cards enabled by multi-level indexing. Further, sentences can be filtered based on classifiers such as oral, written, and technical styles, source, and language difficulty. Additionally sample sentences for terms can be filtered by their inflection and the semantics of a particular definition. Interactivity can be found in the word alignment between the languages as one moves his or her mouse over the words, which can also be clicked on for deeper exploration. And in addition to traditional text-to-speech, a visual representation of a human language tutor pronouncing each sentence is also included. Sample sentences between two similar words can be displayed side-by-side in a tabbed

(a) A screenshot of the definition and sample sentence areas of a *Engkoo* result page.

(b) A screenshot of samples sentences for the POS-wildcard query "v. tv" (meaning "verb TV").

(c) A screenshot of machine translation integrated into the dictionary experience, where the top pane shows results of machine translation while the bottom pane displays example sentences mined from the web.

Figure 2: Three scenarios of *Engkoo*.

user interface to easily expose the subtleties between usages.

In the example seen in Figure 2(b), a user has searched for the collocation verb+TV, represented by the query "v. TV" to find commonly used verbs describing actions for the noun "TV." In the results, we find fresh and authentic sample sentences mined from the web, the first of which contains "watch TV," the most common collocation, as the top result. Additionally, the corresponding keyword in Chinese is automatically highlighted using statistical alignment techniques.

Machine Translation. For many users, the difference between a machine translation (MT) system and a translation dictionary are not entirely clear. In *Engkoo*, if a term or phrase is out-of-vocabulary, a MT result is dynamically returned. For shorter MT queries, sample sentences might also be returned as one can see in Figure 2(c) which expands the search and also raises confidence in a translation as one can observe it used on the web. Like the sample sentences, word alignment is also exposed on the machine translation. As the alignment naturally serves as a word breaker, users can click the selection for a lookup which would open a new tab with the definition. This is especially useful in cases where a user might want to find alternatives to a particular part of a translation. Note that the seemingly single line dictionary search box is also adapted to MT behavior, allowing users to paste in multi-line text as it can detect and unfold itself to a larger text area as needed.

4 Conclusions and Future work

We have presented *Engkoo*, a novel online translation system which uniquely unifies the scenarios of dictionary, machine translation, and language learning. The features of the offering are based on an ever-expanding data set derived from state-of-the-art web mining and NLP techniques. The contribution of the work is a complete software system that maximizes the web's pedagogical potential by exploiting its massive language resources. Direct user feedback and implicit log data suggest that the service is effective for both translation utility and language learning, with advantages over existing services. In future work, we are examining extracting language knowledge from the real-time web for translation in news scenarios. Additionally, we are actively mining other language pairs to build a multi-language learning system.

Acknowledgments

We thank Cheng Niu, Dongdong Zhang, Frank Soong, Gang Chen, Henry Li, Hao Wei, Kan Wang, Long Jiang, Lijuan Wang, Mu Li, Tantan Feng, Weijiang Xu and Yuki Arase for their valuable contributions to this paper, and the anonymous reviewers for their valuable comments.

References

Long Jiang, Ming Zhou, Lee-Feng Chien, and Cheng Niu. 2007. Named entity translation with web mining and transliteration. In *IJCAI*, pages 1629–1634.

Gonglue Jiang, Chen Zhao, Matthew R. Scott, and Fang Zou. 2009a. Combinable tabs: An interactive method of information comparison using a combinable tabbed document interface. In *INTERACT*, pages 432–435.

Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009b. Mining bilingual data from the web with adaptively learnt patterns. In *ACL/AFNLP*, pages 870–878.

Tim Johns. 1991. From printout to handout: grammar and vocabulary teaching in the context of data driven learning. *Special issue of ELR Journal*, pages 27–45.

Mu Li, Nan Duan, Dongdong Zhang, Chi-Ho Li, and Ming Zhou. 2009. Collaborative decoding: Partial hypothesis re-ranking using translation consensus between decoders. In *ACL/AFNLP*, pages 585–592.

Xiaohua Liu and Ming Zhou. 2010. Evaluating the quality of web-mined bilingual sentences using multiple linguistic features. In *IALP*, pages 281–284.

Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller, and Ming Zhou. 2010. Srl-based verb selection for esl. In *EMNLP*, pages 1068–1076.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *ACL*.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *ACL*, pages 489–496.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *ICSLP*, volume 2, pages 901–904.

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou, Zhongyang Xiong, John Lee, and Chin-Yew Lin. 2007. Detecting erroneous sentences using automatically mined sequential patterns. In *ACL*.