Evaluating the Quality of Web-Mined Bilingual Sentence Pairs

Xiaohua Liu[1,2], Ming Zhou[2]

[1] School of Computer Science and Technology
Harbin Institute of Technology,Harbin,150001,China
[2] Microsoft Research Asia,Beijing, 100190,China
{xiaoliu,mingzhou}@microsoft.com

Abstract

We come up with the problem of evaluating the quality of bilingual sentence pairs mined from the web, which is critical for a wide range of applications such as statistical machine translation (SMT) and English as Second Language (ESL) learning. To address this problem, we propose a novel method that integrates multiple linguistic features related to spelling, grammar, alignment, and particularly the sentence type feature that indicates if a sentence can be parsed by the Link Grammar Parser (LGP). Promising results are achieved on a bilingual corpus of about 6 million English-Chinese sentences mined from the web, indicating the effectiveness of our proposed method.

Keywords

Linguistic quality evaluation, bilingual sentence pairs, classification; web mining, linguistic features, sentence type, link grammar parser.

1 Introduction

The automatic acquisition of parallel corpora from the web has played an important role in many applications, including cross-language information retrieval, SMT, and thesaurus construction for ESL learners. Since the quality of mined parallel sentences is critical for such applications, evaluating their quality is thus very meaningful. As a preliminary study, we randomly sampled a bilingual corpus of about 6 million English-Chinese sentences mined from the web, and found that 32% of them are unacceptable due to quality issues. Obviously, manually checking every mined pair is prohibitively impractical. Therefore, an automatic evaluation method is desirable. However mechanical quality evaluation of web mined bilingual corpora is a challenging task, mainly due to the fact that the quality depends on various factors such as grammar correctness, fluentness, word usage correctness, which should be jointly considered.

To tackle this challenge, some researchers(Zhao and Vogel, 2002; Utiyama and Isahara, 2003) use sentence source as the main evidence, and restrict their mining scope to the trusted, reputable, or controlled sites, for example, authoritative bilingual news sites. Other researchers (Liu et al., 2010; Sun et al., 2007; Gamon et al., 2008; Yi et al., 2008; Turner and Charniak, 2007; Fossati and Eugenio, 2007; Izumi et al., 2003; Brockett et al., 2006; Shi and Zhou, 2005; Thurmair, 1990) develop rule-based, statistics-based, or parsing-based filters to detect erroneous sentences automatically, which can be used to filter low quality sentence

pairs. Nevertheless, their work mainly focused on detecting specific errors, and therefore cannot work well for diversified errors that occur in an open domain, like the web.

We propose to use multiple linguistic features to distinguish high-quality bilingual sentences from low-quality ones, while not targeting particular error types. By randomly sampling the 6 million bilingual sentence pairs mined from the web, we observe that 91% of the quality issues are related to English spelling, English grammar, and translational equivalence, while only a few are related to the Chinese part. Inspired by this observation, we mainly consider features regarding English spelling, English grammar, and translational equivalence. We also observe that nearly half of the high-quality sentences are titles or phrases in technical or other specialized domains, which are grammatically correct but cannot be parsed correctly by LGP (Sleator and Temperley, 1993). They would often be judged as low quality sentences in terms of only grammar related features. Therefore, we introduce the sentence type binary feature, which values 1 if the English sentence can be successfully parsed by LGP and 0 otherwise. Table 1 gives some examples of sentences that can and cannot be successfully parsed by LGP, respectively. By considering sentence type and other features, we can identify high quality titles and phrases, even though they do not obey the link grammar.

Our contribution can be summarized as follows: 1) we come up with the problem of evaluating web mined bilingual sentence pairs; 2) we propose to construct a Support Vector Machine (SVM) to integrate various linguistic features, particularly the sentence type feature, to resolve this problem, which is novel; and 3) experimental results show our method is effective.

Our paper is organized as follows. In the next section, we introduce related work. In Section 3, we detail our method. In Section 4, we evaluate our method. Finally, Section 5 concludes and presents future work.

2 Related Work

The most related work is the research of detecting erroneous sentences, which falls into two categories. The first category makes use of hand-crafted rules, e.g., template rules (Heidorn, 2000) and mal-rules in context-free grammars (Michaud et al., 2000; Bender et al., 2004). These methods have shown to be effective in detecting certain kinds of grammar errors. However, writing high-quality rules is time-consuming and labor-intensive.

The second category uses statistical techniques to detect erroneous sentences. (Chodorow and Leacock, 2000) use an unsupervised to detect grammatical errors by inferring negative evidence from the test of English as a foreign language (TOEFL) administrated by the Educational Testing Service (ETS). The method of (Izumi et al., 2003) aims to detect omission-type and replacement-type errors while transformation-based leaning is employed in (Shi and Zhou, 2005) to learn rules for detecting errors in speech recognition outputs. They also require specification of error tags that can indicate the specific errors and their corrections in the training corpus. The phrasal Statistical Machine Translation (SMT) technique has been employed to identify and correct writing errors (Brockett et al., 2006). This method depends on collection of a large number of parallel corpora (pairs of erroneous sentences and their corrections) and its performance depends on SMT techniques that are not yet mature. The work in (Nagata et al., 2006) focuses on a single type of error, namely mass vs. count nouns. Most recently, Liu et al. (2010) propose to use high level linguistic information such as semantic role labeling to detect and correct verb selection errors. Unlike this line of research focusing on detection errors in English text, our work aims to evaluate the quality of bilingual sentence pairs mined from the web.

There are also studies on automatic essay scoring at document-level, for example, E-rater[1] developed by the Educational Testing Service (ETS), and Intelligent Essay Assessor[2]. However, evaluation criteria for documents are different from those for sentences. A document is evaluated by its organization, topic, diversity of vocabulary, and grammar while a sentence is done by grammar, sentence structure, and lexical choice. Different from these works, our work focuses on bilingual sentence pair, not a whole document.

Another related work is Machine Translation (MT) evaluation. Classification models are employed in (Gamon et al., 2005) to evaluate the well-formedness of machine translation outputs. Similarly, we adopt a classifying model, but using different features.

3 Our Method

Our quality evaluation framework consists of two layers, as illustrated in Figure 1. The first layer extracts linguistic features and constructs a unified feature vector to characterize different dimensions of sentence quality, i.e., features w.r.t. English spelling, grammar and alignment. The second layer is a SVM that integrates these linguistic features to detect high-quality bilingual sentences. The following subsections detail each component.

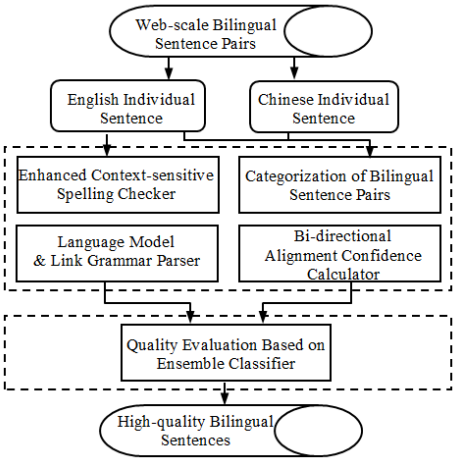

Figure 1. Quality evaluation framework for Bilingual sentence pairs

3.1 Spell Checker Related Features

A spell checker is developed to detect any misspelled word or phrase for the English part of each bilingual pair. The spell checker integrates dictionary-based and context-sensitive spell checkers. The dictionary-based spell checker is based on a composite dictionary, which includes an ESL word list, Encarta, Microsoft Office, TongYi, WordNet and some common out-of-vocabulary terms mined from the web. The context-sensitive spell checker (Fossati

[1] http://www.ets.org/erater/about
[2] http://www.knowledge-technologies.com/prodIEA.shtml

and Eugenio, 2007) uses a mixed trigrams language model for discrimination of confusable words. Finally, the number of misspellings is utilized to describe the spelling attributes.

3.2 Grammar Related Features

Grammar related features are used to represent grammatical correctness of the English sentence. Those features are derived from language model, link grammar parsing result, and sentence type, respectively.

1) Language Model: The language model is a 5-gram model (Gamon et al., 2008) trained on the English Gigaword corpus (LDC2005T12). Kneser-Ney smoothing (Kneser and Ney, 1995) is used to preserve as much context information as possible. The resulting language model score for each English sentence is used as a feature, which indicates the fluentness. Considering that a sentence's language model score decreases as its length increases, we introduce sentence length as a feature as well, which is a common practice in SMT research.

2) Link Grammar Parser: Although not intended for English grammar checking, LGP can provide good evidence of whether or not the English is grammatical via the linkage information generated by it. We apply LGP to the English sentences to get the parsing labels and its linkage information. Here the parsing labels come from a predefined set of link types, and the linkage information gives us the parsing cost vector including the number of unused words, etc. The number of unlinked words is used as the parsing related feature.

3) Sentence Type: Two heuristic rules are designed to decide if an English sentence is parsing consistent or not: if an English sentence has any unlinked word or has less than N words (which is experimentally set to 5), it is parsing consistent; otherwise, not.

3.3 Translational Equivalence Related Features

Translational equivalence measures how well a Chinese and an English sentence can be translated from one to the other. This equivalence is computed in the following way. First we use a comprehensive bilingual dictionary (ESL, Encarta, Office and TongYi, etc.) to do word alignment. Next we collect the number of aligned words in each bilingual sentence and divide them by sentence length to get two normalized alignment confidence scores, which are packed into the linguistic feature vector.

3.4 Model

A SVM model is utilized to integrate all linguistic features. The LibSVM tool (Chang and Lin, 2001) is used for both training and predication.

4 Experiments

This section describes our experiment settings and analyzes the experimental results on a large subset of about 6 million English-Chinese sentences mined from the Web. Those sentences come from various sources, such as forums, personal blogs, as well as bilingual and monolingual web pages.

4.1 Setting

We first randomly sampled 40,000 sentence pairs from the mined corpus. To reduce human labeling efforts, we conduct the following pre-processing steps: 1) automatically check if a sentence pair can be found in a bilingual sentence/phrase pair database, which is compiled from multi dictionaries. If yes, it is naturally of good quality; 2) automatically check if the English part of every pair has a language model score less than *M*, which is experimentally set to 0.0001. If yes, it is regarded as low quality; 3) run our spell checker for the English part of every pair and automatically label those with more than 5 spelling errors as low quality pairs. We then ask one bilingual annotator to label the remaining pairs. Finally we get 12,800 low quality pairs. We conduct 5-fold cross-validation and also a random evaluation, where additional 1,000 bilingual sentences not in the 40,000 pairs are randomly selected from the full corpus and then labeled. Both macro-precision and macro-recall are used as evaluation metrics. Microsoft Word 2007 is used as the baseline.

We use precision *P* and recall *R* to evaluate the performance, as defined below:

$$P = \frac{TP}{TP + FP} \quad , \quad R = \frac{TP}{TP + FN} .$$

Here *TP*, *FP*, *FN* represent the number of true positive, false positive and false negative, respectively.

4.2 Experiments for Determining Sentence Type

By using the two heuristic rules described in Section 2.1, we get about 8,500 parsing non-consistent pairs. A manually check reveals that about 59% of them are actually acceptable in terms of quality. Typical examples of two sentence types are shown in Table 1 in the next page.

4.3 Experiments for Evaluating Sentence Quality

Table 2 displays the performances with different linguistic feature set. S and T denote spelling and translational features, respectively. LM, LGP and ST stand for language model score, LGP related features and sentence type feature, respectively. Different kernel functions are examined and linear kernel gives the best performance after empirical studies. Other settings are default.

Table 2 shows that the sentence type feature plays an important role in that: 1) it can recognize parsing non-consistent sentences as sentences of good quality, which are often wrongly labeled as low quality when only using language model score (LM) and link grammar parsing features (LGP), thus improving the recall; 2) and it improves the precision as well partially owing to its capability of identifying some low quality pairs that are hard to be detected using other features, i.e., parsing non-consistent sentences with reasonable language model scores. We also observe that the random evaluation result is consistent with the cross-validation result, suggesting the good generalization ability of our method.

Sentence type	Typical examples
Parsing consistent sentences	1. What do you desire me to do? 你想要我做什么? 2. Many straws may bind an elephant. 草多可缚象。 3. You may go in now. 您可以进去了。 4. Be at sixes and sevens 乱七八糟 5. Do you accept credit cards 你收信用卡吗
Parsing non-consistent sentences	1. State Natural Science Funds Commission 国家自然科学基金委员会 2. Fig 10 Diagram of Component Interaction In CC 表 10 控制中心交互组件图 3. The Weight Of The Wind 风之悲歌 4. Rule Mining Based on Rough Set 基于粗糙集的规则的挖掘 5. Theory of hard and soft acids and bases 关于软硬酸碱理论

Table 1. Examples of two sentence types

	Cross-validation		Random evaluation	
	P	R	P	R
S+LM	0.5832	0.6092	0.5011	0.6346
S+LM+LGP	0.6019	0.6218	0.5101	0.6301
S+LM+LGP+ST	0.8591	0.8765	0.8100	0.8305
S+LM+LGP+ST+T	0.8826	0.8843	0.8205	0.8296

Table 2. Performance Comparison of Different Features

Table 3 gives some high quality and low quality sentence pairs, all of which are correctly identified by our method.

Category	Typical bilingual sentences	Error
High-quality	1. Every dog has his day. 凡人皆有得意时。 2. All roads lead to Rome. 条条道路通罗马。 3. Let me explain why I was late. 让我解释迟到的理由。	None

	4. I heard someone laughing. 我听见有人在笑。 5. Progress in Microbiology and Immunology 微生物学免疫学进展	
Low-quality	1. **Know is know, no know is no know**. 知之为知之，不知为不知。 2. You **stay a little while** 你呆了一会儿 3. Their **jobless total reached a record high** since 1940. 他们的失业总人数达到 1940 年以来的最高峰。 4. You **stay a little while** 你呆了一会儿 5. What **you look in at**? 你在看什么?	Grammar
	1. **International Union for Conservation of Nature and Natural Resources** 的缩写，是一个 国际组织 ，专职在世界的自然环境保护。	Translation
	1. **d.** This is the question about which we've had so much discussion. 这是我们以讨论了多次的问题。	Noisy symbol
	1. It is a **pitythat** they are not here. 遗憾的是他们没在这里。	Spelling

Table 3. High-quality and low-quality bilingual sentences

We further compare our method with the baseline. Considering that the baseline cannot use alignment related information, to make the comparison fair, we use no translation equivalent related features in our method. A sentence pair counts as low quality if and only if the baseline reports any spelling or grammar errors for its English part. Table 4 displays the comparison result, suggesting our approach significantly outperforms the baseline (p-value <0.001). By sampling the outputs of the baseline, we find that most grammar errors slip through, though nearly all spelling errors are successfully identified. For example, the baseline cannot detect any grammar error for all the grammar incorrect sentence pairs listed in Table 3.

	Cross-validation		Random evaluation	
	P	R	P	R
S+LM+LGP+ST	0.8591	0.8765	0.8100	0.8305
MS Word 2007	0.7531	0.7616	0.7184	0.7220

Table 4. Comparison with MS Word 2007 for quality evaluation on English sentences

We apply the trained model to the 6 million bilingual sentence pairs mined from the web, and find 2,286,164 or 36.1% of all are identified as low-quality pair, from which 1000 pairs are randomly sampled and then labeled to study the error distribution. We find that spelling (41%), grammar (37%) and alignment errors (13%) constitute the majority of errors, while the remaining are the noisy symbols (9%). We also observe that noisy symbols (for example, the last second pair in Table 3) can be easily filtered out during the mining process, and that spelling and alignment errors are largely owing to the mining process (crawling, HTML parsing and bilingual sentence boundary detection). Thus, we believe most serious errors are the grammatical ones.

5 Conclusion and Future Work

This paper raises the problem of evaluating the quality of web-mined bilingual sentences. We propose the use of a SVM to integrate multiple linguistic features, especially a sentence type feature to distinguish high-quality from low-quality bilingual sentence pairs. Experimental results demonstrate the effectiveness of our method. We also experimentally study the distribution of errors in low quality pairs based on a 6 million corpus, which suggests grammar errors require special attention.

In the future, we plan to boost the performance of our method by incorporating other features, such as source and confidence of extraction correctness from the mining component. We also want to apply our method to a real bilingual sentence mining system to filter low quality pairs.

6 Acknowledgments.

We thank the anonymous reviewers for their invaluable comments on an earlier draft of the paper.

7 References

Chris Brockett, William B. Dolan, and Michael Gamon. 2006. *Correcting ESL Errors using Phrasal SMT Techniques*. ACL 2006.

Emily M. Bender, Dan Flickinger, Stephan Oepen, Annemarie Walsh, and Timothy Baldwin. 2004. *Arboretum: Using a precision grammar for grammar checking in call*. In Proc. InSTIL/ICALL Symposium on Computer Assisted Learning.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM : a library for support vector machines* (2001).

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In NAACL.

Davide Fossati and Barbara Eugenio. 2007. *A Mixed Trigrams Approach for Context Sensitive Spell Checking*. CICLing 2007, 623-633.

Michael Gamon, Jianfeng Gao, et al. 2008. *Using Contextual Speller Techniques and Language Modeling for ESL Error Correction*. IJCNLP 2008.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. *Sentence-level MT evaluation without reference translations: beyond language modeling*. In European Association for Machine Translation (EAMT), May.

George E. Heidorn. 2000. *Intelligent Writing Assistance*. *Handbook of Natural Language Processing*. Robert Dale, Hermann Moisi and Harold Somers (ed.). Marcel Dekker.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga et al. 2003. *Automatic Error Detection in the Japanese Learners' English Spoken Data*. In proceedings of ACL 2003.

Reinhard Kneser and Hermann Ney. 1995. *Improved Backing-off for M-gram Language Modeling*. ICASSP 1995, 181–184.

Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller and Ming Zhou. 2010. *SRL-based Verb Selection for ESL*. EMNLP 2010.

Lisa N. Michaud, Kathleen F. McCoy, and Christopher A. Pennington. 2000. *An intelligent tutoring system for deaf learners of written English*. In Proc. 4th International ACM Conference on Assistive Technologies.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. *A feedback-augmented method for detecting errors in the writing of learners of English*. In Proc. ACL.

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou and et al. 2007. *Detecting Erroneous Sentences using Automatically Mined Sequential Patterns*. In ACL 2007.

Yongmei Shi and Lina Zhou. 2005. *Error Detection using Linguistic Features*. HLT/EMNLP 2005.

Daniel D. K. Sleator and Davy Temperley. 1993. *Parsing English with a Link Grammar*. Third International Workshop on Parsing Technologies 1993.

Jenine Turner and Eugene Charniak. 2007. *Language Modeling for Determiner Selection*. In proceedings of NAACL 2007, Short Papers, 177-180.

Gregor Thurmair, *Parsing for Grammar and Style Checking*, 1990, ACL 1990, 365-370.

Xing Yi, Jianfeng Gao and William Dolan. 2008. *A Web-based English Proofing System for ESL Users*. In proceedings of IJCNLP 2008.

Bing Zhao and Stephan Vogel. 2002. *Adaptive Parallel Sentences Mining From Web Bilingual News Collection*. In ICDM 2002.