Detection of Non-native Sentences using Machine-translated Training Data

John Lee
Spoken Language Systems
MIT CSAIL
Cambridge, MA 02139, USA
jsylee@csail.mit.edu

Ming Zhou, Xiaohua Liu
Natural Language Computing Group
Microsoft Research Asia
Beijing, 100080, China
{mingzhou,xiaoliu}@microsoft.com

Abstract

Training statistical models to detect non-native sentences requires a large corpus of non-native writing samples, which is often not readily available. This paper examines the extent to which machine-translated (MT) sentences can substitute as training data.

Two tasks are examined. For the native vs non-native *classification* task, non-native training data yields better performance; for the *ranking* task, however, models trained with a large, publicly available set of MT data perform as well as those trained with non-native data.

1 Introduction

For non-native speakers writing in a foreign language, feedback from native speakers is indispensable. While humans are likely to provide higher-quality feedback, a computer system can offer better availability and privacy. A system that can distinguish *non-native* ("ill-formed") English sentences from *native* ("well-formed") ones would provide valuable assistance in improving their writing.

Classifying a sentence into discrete categories can be difficult: a sentence that seems fluent to one judge might not be good enough to another. An alternative is to rank sentences by their relative fluency. This would be useful when a non-native speaker is unsure which one of several possible ways of writing a sentence is the best.

We therefore formulate two tasks on this problem. The **classification** task gives one sentence to the system, and asks whether it is native or non-native. The **ranking** task submits sentences with the same intended meaning, and asks which one is best.

To tackle these tasks, hand-crafting formal rules would be daunting. Statistical methods, however, require a large corpus of non-native writing samples, which can be difficult to compile. Since machine-translated (MT) sentences are readily available in abundance, we wish to address the question of whether they can substitute as training data.

The next section provides background on related research. Sections 3 and 4 describe our experiments, followed by conclusions and future directions.

2 Related Research

Previous research has paid little attention to ranking sentences by fluency. As for classification, one line of research in MT evaluation is to evaluate the fluency of an output sentence without its reference translations, such as in (Corston-Oliver et al., 2001) and (Gamon et al., 2005). Our task here is similar, but is applied on non-native sentences, arguably more challenging than MT output.

Evaluation of non-native writing has encompassed both the document and sentence levels. At the document level, automatic essay scorers, such as (Burstein et al., 2004) and (Ishioka and Kameda, 2006), can provide holistic scores that correlate well with those of human judges.

At the sentence level, which is the focus of this paper, previous work follows two trends. Some researchers explicitly focus on individual classes of er-

Proceedings of NAACL HLT 2007, Companion Volume, pages 93–96,
Rochester, NY, April 2007.

rors, e.g., mass vs count nouns in (Brockett et al., 2006) and (Nagata et al., 2006). Others implicitly do so with hand-crafted rules, via templates (Heidorn, 2000) or mal-rules in context-free grammars, such as (Michaud et al., 2000) and (Bender et al., 2004).

Typically, however, non-native writing exhibits a wide variety of errors, in grammar, style and word collocations. In this research, we allow unrestricted classes of errors[1], and in this regard our goal is closest to that of (Tomokiyo and Jones, 2001). However, they focus on non-native speech, and assume the availability of non-native training data.

3 Experimental Set-Up

3.1 Data

Our data consists of pairs of English sentences, one native and the other non-native, with the same "intended meaning". In our MT data (`MT`), both sentences are translated, by machine or human, from the same sentence in a foreign language. In our non-native data (`JLE`), the non-native sentence has been edited by a native speaker[2]. Table 1 gives some examples, and Table 2 presents some statistics.

`MT` (Multiple-Translation Chinese and Multiple-Translation Arabic corpora) English MT output, and human reference translations, of Chinese and Arabic newspaper articles.

`JLE` (Japanese Learners of English Corpus) Transcripts of Japanese examinees in the Standard Speaking Test. False starts and disfluencies were then cleaned up, and grammatical mistakes tagged (Izumi et al., 2003). The speaking style is more formal than spontaneous English, due to the examination setting.

3.2 Machine Learning Framework

SVM-Light (Joachims, 1999), an implementation of Support Vector Machines (SVM), is used for the classification task.

For the ranking task, we utilize the ranking mode of SVM-Light. In this mode, the SVM algorithm is adapted for learning ranking functions, originally used for ranking web pages with respect to a query (Joachims, 2002). In our context, given a set of English sentences with similar semantic content, say $s_1, \ldots, s_n$, and a ranking based on their fluency, the learning algorithm estimates the weights $\vec{w}$ to satisfy the inequalities:

$$\vec{w} \cdot \Phi(s_j) > \vec{w} \cdot \Phi(s_k) \quad (1)$$

where s_j is more fluent than s_k, and where Φ maps a sentence to a feature vector. This is in contrast to standard SVMs, which learn a hyperplane boundary between native and non-native sentences from the inequalities:

$$y_i(\vec{w} \cdot \Phi(s_i) + w_0) - 1 \geq 0 \quad (2)$$

where $y_i = \pm 1$ are the labels. Linear kernels are used in our experiments, and the regularization parameter is tuned on the development sets.

3.3 Features

The following features are extracted from each sentence. The first two are real numbers; the rest are indicator functions of the presence of the lexical and/or syntactic properties in question.

`Ent` Entropy[3] from a trigram language model trained on 4.4 million English sentences with the SRILM toolkit (Stolcke, 2002). The trigrams are intended to detect local mistakes.

`Parse` Parse score from Model 2 of the statistical parser (Collins, 1997), normalized by the number of words. We hypothesize that non-native sentences are more likely to receive lower scores.

`Deriv` Parse tree derivations, i.e., from each parent node to its children nodes, such as S → NP VP. Some non-native sentences have plausible N-grams, but have derivations infrequently seen in well-formed sentences, due to their unusual syntactic structures.

`DtNoun` Head word of a base noun phrase, and its determiner, e.g., (*the, markets*) from the human non-native sentence in Table 1. The usage of articles has been found to be the most frequent error class in the `JLE` corpus (Izumi et al., 2003).

[1] Except spelling mistakes, which we consider to be a separate problem that should be dealt with in a pre-processing step.

[2] The nature of the non-native data constrains the ranking to two sentences at a time.

[3] Entropy $H(x)$ is related to perplexity $PP(x)$ by the equation $PP(x) = 2^{H(x)}$.

Type		Sentence
Native	Human	New York and London stock markets went up
Non-native	Human	The stock markets in New York and London were increasing together
	MT	The same step of stock market of London of New York rises

Table 1: *Examples of sentences translated from a Chinese source sentence by a native speaker, by a non-native speaker, and by a machine translation system.*

Data Set	Corpus	# sentences (for classification)			# pairs (for ranking)
		total	native	non-native	
`MT train`	LDC{2002T01, 2003T18, 2006T04}	30075	17508	12567	91795
`MT dev`	LDC2003T17 (*Zaobao* only)	1995	1328	667	2668
`MT test`	LDC2003T17 (*Xinhua* only)	3255	2184	1071	4284
`JLE train`	Japanese Learners of English	9848	4924	4924	4924
`JLE dev`		1000	500	500	500
`JLE test`		1000	500	500	500

Table 2: *Data sets used in this paper.*

`Colloc` An in-house dependency parser extracts five types of word dependencies[4]: subject-verb, verb-object, adjective-noun, verb-adverb and preposition-object. For the human non-native sentence in Table 1, the unusual subject-verb collocation "*market increase*" is a useful clue in this otherwise well-formed sentence.

4 Analysis

4.1 An Upper Bound

To gauge the performance upper bound, we first attempt to classify and rank the MT test data, which should be less challenging than non-native data. After training the SVM on `MT train`, classification accuracy on `MT test` improves with the addition of each feature, culminating at 89.24% with all five features. This result compares favorably with the state-of-the-art[5]. Ranking performance reaches 96.73% with all five features.

We now turn our attention to non-native test data, and contrast the performance on `JLE test` using models trained by MT data (`MT train`), and by non-native data (`JLE train`).

[4]Proper nouns and numbers are replaced with special symbols. The words are further stemmed using Porter's Stemmer.

[5]Direct comparison is impossible since the corpora were different. (Corston-Oliver et al., 2001) reports 82.89% accuracy on English software manuals and online help documents, and (Gamon et al., 2005) reports 77.59% on French technical documents.

Test Set: `JLE test`	Train Set	
	`MT train`	`JLE train`
`Ent+ Parse`	57.2 (+) 48.6 (-) 65.8	57.7 (+) 70.6 (-) 44.8
`+Deriv`	58.4 (+) 54.6 (-) 62.2	64.7 (+)72.2 (-) 57.2
`+DtNoun`	**59.0** (+) 57.6 (-) 60.4	**66.4** (+) 72.8 (-) 60.0
`+Colloc`	58.6 (+) 54.2 (-) 63.2	65.9 (+) 72.6 (-) 59.2

Table 3: *Classification accuracy on `JLE test`. (-) indicates accuracy on non-native sentences, and (+) indicates accuracy on native sentences. The overall accuracy is their average.*

4.2 Classification

As shown in Table 3, classification accuracy on `JLE test` is higher with the `JLE train` set (66.4%) than with the larger `MT train` set (59.0%). The SVM trained on `MT train` consistently misclassifies more native sentences than non-native ones. One reason might be that speech transcripts have a less formal style than written news sentences. Transcripts of even good conversational English do not always resemble sentences in the news domain.

4.3 Ranking

In the ranking task, the relative performance between MT and non-native training data is reversed.

Test Set: JLE test	Train Set	
	MT train	JLE train
Ent+Parse	72.8	71.4
+Deriv	73.4	73.6
+DtNoun	75.4	73.8
+Colloc	**76.2**	**74.6**

Table 4: *Ranking accuracy on JLE test.*

As shown in Table 4, models trained on MT train yield higher ranking accuracy (76.2%) than those trained on JLE train (74.6%). This indicates that MT training data can generalize well enough to perform better than a non-native training corpus of size up to 10000.

The contrast between the classification and ranking results suggests that train/test data mismatch is less harmful for the latter task. Weights trained on the classification inequalities in (2) and on the ranking inequalities in (1) both try to separate native and MT sentences maximally. The absolute boundary learned in (2) is inherently specific to the nature of the training sentences, as we have seen in §4.2. In comparison, the relative scores learned from (1) have a better chance to carry over to other domains, as long as some gap still exists between the scores of the native and non-native sentences.

5 Conclusions & Future Work

We explored two tasks in sentence-level fluency evaluation: ranking and classifying native vs. non-native sentences. In an SVM framework, we examined how well MT data can replace non-native data in training.

For the classification task, training with MT data is less effective than with non-native data. However, for the ranking task, models trained on publicly available MT data generalize well, performing as well as those trained with a non-native corpus of size 10000.

In the future, we would like to search for more salient features through a careful study of non-native errors, using error-tagged corpora such as (Izumi et al., 2003). We also plan to explore techniques for combining large MT training corpora and smaller non-native training corpora. Our ultimate goal is to identify the errors in the non-native sentences and propose corrections.

References

E. Bender, D. Flickinger, S. Oepen, A. Walsh, and T. Baldwin. 2004. Arboretum: Using a Precision Grammar for Grammar Checking in CALL. *Proc. InSTIL/ICALL Symposium on Computer Assisted Learning*.

C. Brockett, W. Dolan, and M. Gamon. 2006. Correcting ESL Errors using Phrasal SMT Techniques. *Proc. ACL.*

J. Burstein, M. Chodorow and C. Leacock. 2004. Automated Essay Evaluation: The Criterion online Writing Service. *AI Magazine*, 25(3):27–36.

M. Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. *Proc. ACL.*

S. Corston-Oliver, M. Gamon and C. Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. *Proc. ACL.*

M. Gamon, A. Aue, and M. Smets. 2005. Sentence-Level MT Evaluation without Reference Translations: Beyond Language Modeling. *Proc. EAMT.*

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

T. Ishioka and M. Kameda. 2006. Automated Japanese Essay Scoring System based on Articles Written by Experts. *Proc. ACL.*

E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic Error Detection in the Japanese Learners' English Spoken Data. *Proc. ACL.*

T. Joachims. 1999. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*. B. Schölkopf, C. Burges and A. Smola (ed.), MIT-Press.

T. Joachims. 2002. Optimizing Search Engines using Clickthrough Data. *Proc. SIGKDD.*

L. Michaud, K. McCoy and C. Pennington. 2000. An Intelligent Tutoring System for Deaf Learners of Written English. *Proc. 4th International ACM Conference on Assistive Technologies*.

R. Nagata, A. Kawai, K. Morihiro, and N. Isu. 2006. A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. *Proc. ACL.*

A. Stolcke. 2002. SRILM — An Extensible Language Modeling Toolkit *Proc. ICSLP.*

L. Tomokiyo and R. Jones. 2001. You're not from 'round here, are you? Naïve Bayes Detection of Non-native Utterance Text. *Proc. NAACL.*