NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Bilingual sentiment consistency for statistical machine translation

Chen, Boxing; Zhu, Xiaodan

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26-30 2014, pp. 607-615, 2014-04-30

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=d23168f2-8ee4-4559-8a93-bf5b228c3660
https://publications-cnrc.canada.ca/fra/voir/objet/?id=d23168f2-8ee4-4559-8a93-bf5b228c3660

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

Bilingual Sentiment Consistency for Statistical Machine Translation

Boxing Chen and **Xiaodan Zhu**
National Research Council Canada
1200 Montreal Road, Ottawa, Canada, K1A 0R6
{Boxing.Chen, Xiaodan.Zhu}@nrc-cnrc.gc.ca

Abstract

In this paper, we explore bilingual sentiment knowledge for statistical machine translation (SMT). We propose to explicitly model the consistency of sentiment between the source and target side with a lexicon-based approach. The experiments show that the proposed model significantly improves Chinese-to-English NIST translation over a competitive baseline.

1 Introduction

The expression of *sentiment* is an interesting and integral part of human languages. In written text sentiment is conveyed by s*enses* and in speech also via *prosody*. Sentiment is associated with both *evaluative* (positive or negative) and *potency* (degree of sentiment) — involving two of the three major semantic differential categories identified by Osgood et al. (1957).

Automatically analyzing the sentiment of monolingual text has attracted a large bulk of research, which includes, but is not limited to, the early exploration of (Turney, 2002; Pang et al., 2002; Hatzivassiloglou & McKeown, 1997). Since then, research has involved a variety of approaches and been conducted on various type of data, e.g., product reviews, news, blogs, and the more recent social media text.

As sentiment has been an important concern in monolingual settings, better translation of such information between languages could be of interest to help better cross language barriers, particularly for sentiment-abundant data. Even when we randomly sampled a subset of sentence pairs from the NIST Open MT[1] training data, we found that about 48.2% pairs contain at least one sentiment word on both sides, and 22.4% pairs contain at least one intensifier word on both sides, which suggests a non-trivial percent of sentences may potentially involve sentiment in some degree[2].

# snt. pairs	% snt. with sentiment words	% snt. with intensifiers
103,369	48.2%	22.4%

Table 1. Percentages of sentence pairs that contain sentiment words on both sides or intensifiers[3] on both sides.

One expects that sentiment has been implicitly captured in SMT through the statistics learned from parallel corpus, e.g., the phrase tables in a phrase-based system. In this paper, we are interested in explicitly modeling sentiment knowledge for translation. We propose a lexicon-based approach that examines the consistency of bilingual subjectivity, sentiment polarity, intensity, and negation. The experiments show that the proposed approach improves the NIST Chinese-to-English translation over a strong baseline.

In general, we hope this line of work will help achieve better MT quality, especially for data with more abundant sentiment, such as social media text.

2 Related Work

Sentiment analysis and lexicon-based approaches Research on monolingual sentiment analysis can be found under different names such as *opinion*, *stance*, *appraisal*, and *semantic orientation,* among others. The overall goal is to label a span of text either as positive, negative, or neutral — sometimes the strength of sentiment is a concern too.

[1] http://www.nist.gov/speech/tests/mt

[2] The numbers give a rough idea of sentiment coverage; it would be more ideal if the estimation could be conducted on senses instead of words, which, however, requires reliable sense labeling and is not available at this stage. Also, according to our human evaluation on a smaller dataset, two thirds of such potentially sentimental sentences do convey sentiment.

[3] The sentiment and intensifier lexicons used to acquire these numbers are discussed in Section 3.2.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 607–615,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

The granularities of text have spanned from words and phrases to passages and documents.

Sentiment analysis has been approached mainly as an unsupervised or supervised problem, although the middle ground, semi-supervised approaches, exists. In this paper, we take a lexicon-based, unsupervised approach to considering sentiment consistency for translation, although the translation system itself is supervised. The advantages of such an approach have been discussed in (Taboada et al., 2011). Briefly, it is good at capturing the basic sentiment expressions common to different domains, and certainly it requires no bilingual sentiment-annotated data for our study. It suits our purpose here of exploring the basic role of sentiment for translation. Also, such a method has been reported to achieve a good cross-domain performance (Taboada et al., 2011) comparable with that of other state-of-the-art models.

Translation for sentiment analysis A very interesting line of research has leveraged labeled data in a resource-rich language (e.g., English) to help sentiment analysis in a resource-poorer language. This includes the idea of constructing sentiment lexicons automatically by using a translation dictionary (Mihalcea et al., 2007), as well as the idea of utilizing parallel corpora or automatically translated documents to incorporate sentiment-labeled data from different languages (Wan, 2009; Mihalcea et al., 2007).

Our concern here is different — instead of utilizing translation for sentiment analysis; we are interested in the SMT quality itself, by modeling bilingual sentiment in translation. As mentioned above, while we expect that statistics learned from parallel corpora have implicitly captured sentiment in some degree, we are curious if better modeling is possible.

Considering semantic similarity in translation The literature has included interesting ideas of incorporating different types of semantic knowledge for SMT. A main stream of recent efforts have been leveraging semantic roles (Wu and Fung, 2009; Liu and Gildea, 2010; Li et al., 2013) to improve translation, e.g., through improving reordering. Also, Chen et al. (2010) have leveraged sense similarity between source and target side as additional features. In this work, we view a different dimension, i.e., semantic orientation, and show that incorporating such knowledge improves the translation performance. We hope this work would add more evidences to the existing literature of leveraging semantics for SMT, and shed some light on further exploration of semantic consistency, e.g., examining other semantic differential factors.

3 Problem & Approach

3.1 Consistency of sentiment

Ideally, sentiment should be properly preserved in high-quality translation. An interesting study conducted by Mihalcea et al. (2007) suggests that in most cases the sentence-level subjectivity is preserved by human translators. In their experiments, one English and two Romanian native speakers were asked to independently annotate the sentiment of English-Romanian sentence pairs from the SemCor corpus (Miller et al., 1993), a balanced corpus covering a number of topics in sports, politics, fashion, education, and others. These human subjects were restricted to only access and annotate the sentiment of their native-language side of sentence pairs. The sentiment consistency was observed by examining the annotation on both sides.

Automatic translation should conform to such a consistency too, which could be of interest for many applications, particularly for sentiment-abundant data. On the other hand, if consistency is not preserved for some reason, e.g., alignment noise, enforcing consistency may help improve the translation performance. In this paper, we explore bilingual sentiment consistency for translation.

3.2 Lexicon-based bilingual sentiment analysis

To capture bilingual sentiment consistency, we use a lexicon-based approach to sentiment analysis. Based on this, we design four groups of features to represent the consistency.

The basic idea of the lexicon-based approach is first identifying the sentiment words, intensifiers, and negation words with lexicons, and then calculating the sentiment value using manually designed formulas. To this end, we adapted the approaches of (Taboada et al., 2011) and (Zhang et al., 2012) so as to use the same formulas to analyze the sentiment on both the source and the target side.

The English and Chinese sentiment lexicons we used are from (Wilson et al. 2005) and (Xu and Lin, 2007), respectively. We further use 75 English in-

tensifiers listed in (Benzinger, 1971; page 171) and 81 Chinese intensifiers from (Zhang et al., 2012). We use 17 English and 13 Chinese negation words.

Similar to (Taboada et al., 2011) and (Zhang et al., 2012), we assigned a numerical score to each sentiment word, intensifier, and negation word. More specifically, one of the five values: -0.8, -0.4, 0, 0.4, and 0.8, was assigned to each sentiment word in both the source and target sentiment lexicons, according to the strength information annotated in these lexicons. The scores indicate the strength of sentiment. Table 2 lists some examples. Similarly, one of the 4 values, i.e., -0.5, 0.5, 0.7 and 0.9, was manually assigned to each intensifier word, and a -0.8 or -0.6 to the negation words. All these scores will be used below to modify and shift the sentiment value of a sentiment unit.

Sentiment words	Intensifiers	Negation words
impressive (0.8)	extremely (0.9)	not (-0.8)
good (0.4)	very (0.7)	rarely (-0.6)
actually (0.0)	pretty (0.5)	
worn (-0.4)	slightly (-0.5)	
depressing (-0.8)		

Table 2: Examples of sentiment words and their sentiment strength; intensifiers and their modify rate; negation words and their negation degree.

Each sentiment word and its modifiers (negation words and intensifiers) form a sentiment unit. We first found all sentiment units by identifying sentiment words with the sentiment lexicons and their modifiers with the corresponding lexicon in a 7-word window. Then, for different patterns of sentiment unit, we calculated the sentiment values using the formulas listed in Table 3, where these formulas are adapted from (Taboada et al., 2011) and (Zhang et al., 2012) so as to be applied to both languages.

Sen. unit	Sen. value formula	Example	Sen. value
w_s	$S(w_s)$	good	0.40
$w_n w_s$	$D(w_n)S(w_s)$	not good	-0.32
$w_i w_s$	$(1+R(w_i))S(w_s)$	very good	0.68
$w_n w_i w_s$	$(1+ D(w_n)R(w_i))S(w_s)$	not very good	0.176
$w_i w_n w_s$	$D(w_n)(1+R(w_i))S(w_s)$	very not good[4]	-0.544

Table 3: Heuristics used to compute the lexicon-based sentiment values for different types of sentiment units.

[4] The expression "very not good" is ungrammatical in English. However, in Chinese, it is possible to have this kind of expression, such as "很不漂亮", whose transliteration is "very not beautiful", meaning "very ugly".

For notation, $S(w_s)$ stands for the strength of sentiment word w_s, $R(w_i)$ is degree of the intensifier word w_i, and $D(w_n)$ is the negation degree of the negation word w_n.

Above, we have calculated the lexicon based sentiment value (LSV) for any given unit u_i, and we call it $lsv(u_i)$ below. If a sentence or phrase s contains multiple sentiment units, its *lsv*-score is a merge of the individual *lsv*-scores of all its sentiment units:

$$lsv(s) = merg_1^N(basis(lsv(u_i))) \quad (1)$$

where the function *basis*(.) is a normalization function that performs on each $lsv(u_i)$. For example, the *basis*(.) function could be a standard *sign* function that just examines if a sentiment unit is positive or negative, or simply an identity function (using the *lsv*-scores directly). The *merg*(.) is a function that merge the *lsv*-scores of individual sentiment units, which may take several different forms below in our feature design. For example, it can be a *mean* function to take the average of the sentiment units' *lsv*-scores, or a logic OR function to examine if a sentence or phrase contains positive or negative units (depending on the basis function). It can also be a linear function that gives different weights to different units according to further knowledge, e.g., syntactic information. In this paper, we only leverage the basic, surface-level analysis[5].

In brief, our model here can be thought of as a unification and simplification of both (Taboada et al., 2011) and (Zhang et al., 2012), for our bilingual task. We suspect that better sentiment modeling may further improve the general translation performance or the quality of sentiment in translation. We will discuss some directions we think interesting in the future work section.

3.3 Incorporating sentiment consistency into phrase-based SMT

In this paper, we focus on exploring sentiment consistency for phrase-based SMT. However, the approach might be used in other translation framework. For example, consistency may be considered in the variables used in hierarchical translation rules (Chiang, 2005).

[5] Note that when sentiment-annotated training data are available, *merg*(.) can be trained, e.g., if assuming it to be the widely-used (log-) linear form.

We will examine the role of sentiment consistency in two ways: designing features for the translation model and using them for re-ranking. Before discussing the details of our features, we briefly recap phrase-based SMT for completeness.

Given a source sentence *f*, the goal of statistical machine translation is to select a target language string *e* which maximizes the posterior probability P(*e*|*f*). In a phrase-based SMT system, the translation unit is the phrases, where a "phrase" is a sequence of words. Phrase-based statistical machine translation systems are usually modeled through a log-linear framework (Och and Ney, 2002) by introducing the hidden word alignment variable *a* (Brown et al., 1993).

$$\tilde{e}^* = \arg\max_{e,a} \left(\sum_{m=1}^{M} \lambda_m H_m(\tilde{e}, \tilde{f}, a)\right) \quad (2)$$

where $\tilde{e}$ is a string of phrases in the target language, $\tilde{f}$ is the source language string, $H_m(\tilde{e}, \tilde{f}, a)$ are feature functions, and weights λ_m are typically optimized to maximize the scoring function (Och, 2003).

3.4 Feature design

In Section 3.2 above, we have discussed our lexicon-based approach, which leverages lexicon-based sentiment consistency. Below, we describe the specific features we designed for our experiments. For a phrase pair ($\tilde{f}, \tilde{e}$) or a sentence pair (*f*, *e*)[6], we propose the following four groups of consistency features.

Subjectivity The first group of features is designed to check the subjectivity of a phrase or a sentence pair (*f*, *e*). This set of features examines if the source or target side contains sentiment units. As the name suggests, these features only capture if *subjectivity* exists, but not if a sentiment is positive, negative, or neutral. We include four binary features that are triggered in the following conditions—satisfaction of each condition gives the corresponding feature a value of 1 and otherwise 0.

- F1: if neither side of the pair (f, e) contains at least one sentiment unit;
- F2: if only one side contains sentiment units;
- F3: if the source side contains sentiment units;
- F4: if the target side contains sentiment units.

Sentiment polarity The second group of features check the sentiment polarity. These features are still binary; they check if the polarities of the source and target side are the same.

- F5: if the two sides of the pair (f, e) have the same polarity;
- F6: if at least one side has a neutral sentiment;
- F7: if the polarity is opposite on the two sides, i.e., one is positive and one is negative.

Note that examining the polarity on each side can be regarded as a special case of applying Equation 1 above. For example, examining the positive sentiment corresponds to using an indicator function as the *basis* function: it takes a value of 1 if the *lsv*-score of a sentiment unit is positive or 0 otherwise, while the *merge* function is the logic OR function. The subjectivity features above can also be thought of similarly.

Sentiment intensity The third group of features is designed to capture the degree of sentiment and these features are numerical. We designed two types of features in this group.

Feature *F8* measures the difference of the LSV scores on the two sides. As shown in Equation (3), we use a *mean* function[7] as our *merge* function when computing the *lsv*-scores with Equation (1), where the *basis* function is simply the identity function.

$$lsv_1(s) = \frac{1}{n}\sum_{i=0}^{n} lsv(u_i) \quad (3)$$

Feature *F9, F10,* and *F11* are the second type in this group of features, which compute the ratio of sentiment units on each side and examine their difference.

- F8: $H_8(f,e) = |lsv_1(f) - lsv_1(e)|$
- F9: $H_9(f,e) = |lsv_+(f) - lsv_+(e)|$

[6] For simplicity, we hereafter use the same notation (*f*, *e*) to represent both a phrase pair and a sentence pair, when no confusion arises.

[7] We studied several different options but found the *average* function is better than others for our translation task here, e.g., better than giving more weight to the last unit.

- F10: $H_{10}(f,e) = |lsv_{-}(f) - lsv_{-}(e)|$
- F11: $H_{11}(f,e) = |lsv_{+-}(f) - lsv_{+-}(e)|$

$lsv_{+}(.)$ calculates the ratio of a positive sentiment units in a phrase or a sentence, i.e., the number of positive sentiment units divided by the total number of words of the phrase or the sentence. It corresponds to a special form of Equation 1, in which the *basis* function is an indicator function as discussed above, and the *merge* function adds up all the counts and normalizes the sum by the length of the phrase or the sentence concerned. Similarly, $lsv_{-}(.)$ calculates the ratio of negative units and $lsv_{+-}(.)$ calculates that for both types of units. The length of sentence here means the number of word tokens. We experimented with and without removing stop words when counting them, and found that decision has little impact on the performance. We also used the part-of-speech (POS) information in the sentiment lexicons to help decide if a word is a sentiment word or not, when we extract features; i.e., a word is considered to have sentiment only if its POS tag also matches what is specified in the lexicons[8]. Using POS tags, however, did not improve our translation performance.

Negation The fourth group of features checks the consistency of negation words on the source and target side. Note that negation words have already been considered in computing the *lsv*-scores of sentiment units. One motivation is that a negation word may appear far from the sentiment word it modifies, as mentioned in (Taboada et al., 2011) and may be outside the window we used to calculate the *lsv*-score above. The features here additionally check the counts of negation words. This group of features is binary and triggered by the following conditions.

- F12: if neither side of the pair (*f, e*) contain negation words;
- F13: if both sides have an odd number of negation words or both sides have an even number of them;
- F14: if both sides have an odd number of negation words not appearing outside any sentiment units, or if both sides have an even number of such negation words;
- F15: if both sides have an odd number of negation words appearing in all sentiment units, or if both sides have an even number of such negation words.

[8] The Stanford POS tagger (Toutanova et al., 2003) was used to tag phrase and sentence pairs for this purpose.

4 Experiments

4.1 Translation experimental settings

Experiments were carried out with an in-house phrase-based system similar to Moses (Koehn et al., 2007). Each corpus was word-aligned using IBM model 2, HMM, and IBM model 4, and the phrase table was the union of phrase pairs extracted from these separate alignments, with a length limit of 7. The translation model was smoothed in both directions with Kneser-Ney smoothing (Chen et al., 2011). We use the hierarchical lexicalized reordering model (Galley and Manning, 2008), with a distortion limit of 7. Other features include lexical weighting in both directions, word count, a distance-based RM, a 4-gram LM trained on the target side of the parallel data, and a 6-gram English *Gigaword* LM. The system was tuned with batch lattice MIRA (Cherry and Foster, 2012).

We conducted experiments on NIST Chinese-to-English translation task. The training data are from NIST Open MT 2012. All allowed bilingual corpora were used to train the translation model and reordering models. There are about 283M target word tokens. The development (dev) set comprised mainly data from the NIST 2005 test set, and also some balanced-genre web-text from NIST training data. Evaluation was performed on NIST 2006 and 2008, which have 1,664 and 1,357 sentences, 39.7K and 33.7K source words respectively. Four references were provided for all dev and test sets.

4.2 Results

Our evaluation metric is case-insensitive IBM BLEU (Papineni et al., 2002), which performs matching of n-grams up to n = 4; we report BLEU scores on two test sets NIST06 and NIST08. Following (Koehn, 2004), we use the bootstrap resampling test to do significance testing. In Table 4-6, the sign * and ** denote statistically significant gains over the baseline at the $p < 0.05$ and $p < 0.01$ level, respectively.

	NIST06	NIST08	Avg.
Baseline	35.1	28.4	31.7
+feat. group1	35.6**	29.0**	32.3
+feat. group2	35.3*	28.7*	32.0
+feat. group3	35.3	28.7*	32.0
+feat. group4	35.5*	28.8*	32.1
+feat. group1+2	35.8**	29.1**	32.5
+feat. group1+2+3	36.1**	29.3**	32.7
+feat. group1+2+3+4	**36.2****	**29.4****	**32.8**

Table 4: BLEU(%) scores on two original test sets for different feature combinations. The sign * and ** indicate statistically significant gains over the baseline at the $p < 0.05$ and $p < 0.01$ level, respectively.

Table 4 summarizes the results of the baseline and the results of adding each group of features and their combinations. We can see that each individual feature group improves the BLEU scores of the baseline, and most of these gains are significant. Among the feature groups, the largest improvement is associated with the first feature group, i.e., the subjectivity features, which suggests the significant role of modeling the basic subjectivity. Adding more features results in further improvement; the best performance was achieved when using all these sentiment consistency features, where we observed a 1.1 point improvement on the NIST06 set and a 1.0 point improvement on the NIST08 set, which yields an overall improvement of about 1.1 BLEU score.

To further observe the results, we split each of the two (i.e., the NIST06 and NIST08) test sets into three subsets according to the ratio of sentiment words in the reference. We call them low-sen, mid-sen and high-sen subsets, denoting lower, middle, and higher sentiment-word ratios, respectively. The three subsets contain roughly equal number of sentences. Then we merged the two low-sen subsets together, and similarly the two mid-sen and high-sen subsets together, respectively. Each subset has roughly 1007 sentences.

	low-sen	mid-sen	high-sen
baseline	33.4	32.3	29.3
+all feat.	34.4**	33.5**	30.4**
improvement	1.0	1.2	1.1

Table 5: BLEU(%) scores on three sub test sets with different sentiment ratios.

Table 5 shows the performance of baseline and the system with sentiment features (the last system of Table 4) on these subsets. First, we can see that both systems perform worse as the ratio of sentiment words increases. This probably indicates that text with more sentiment is harder to translate than text with less sentiment. Second, it is interesting that the largest improvement is seen on the mid-sen sub-set. The larger improvement on the mid-sen/high-sen subsets than on the low-sen may indicate the usefulness of the proposed features in capturing sentiment information. The lower improvement on high-sen than on mid-sen probably indicates that the high-sen subset is hard anyway and using simple lexicon-level features is not sufficient.

Sentence-level reranking Above, we have incorporated sentiment features into the phrase tables. To further confirm the usefulness of the sentiment consistency features, we explore their role for sentence-level reranking. To this end, we re-rank 1000-best hypotheses for each sentence that were generated with the baseline system. All the sentiment features were recalculated for each hypothesis. We then re-learned the weights for the decoding and sentiment features to select the best hypothesis. The results are shown in Table 6. We can see that sentiment features improve the performance via re-ranking. The improvement is statistically significant, although the absolute improvement is less than that obtained by incorporating the sentiment features in decoding. Not that as widely known, the limited variety of candidates in reranking may confine the improvement that could be achieved. Better models on the sentence level are possible. In addition, we feel that ensuring sentiment and its target to be correctly paired is of interest. Note that we have also combined the last system in Table 4 with the reranking system here; i.e., sentiment consistency was incorporated in both ways, but we did not see further improvement, which suggests that the benefit of the sentiment features has mainly been captured in the phrase tables already.

feature	NIST06	NIST08	Avg.
baseline	35.1	28.4	31.7
+ all feat.	35.4*	28.9**	32.1

Table 6: BLEU(%) scores on two original test sets on sentence-level sentiment features.

Human evaluation We conducted a human evaluation on the output of the baseline and the system that incorporates all the proposed sentiment features (the last system in Table 4). For this purpose, we randomly sampled 250 sentences from the two NIST test sets according to the following conditions. First, the selected sentences should contain at least one sentiment word—in this evaluation, we target the sentences that may convey some sentiment. Second, we do not consider sentences shorter than 5 words or longer than 50 words; or where outputs of the baseline system and the system with sentiment feature were identical. The 250 selected sentences were split into 9 subsets, as we have 9 human evaluators (none of the authors of this paper took part in this experiment). Each subset contains 26 randomly selected sentences, which are 234 sentences in total. The other 16 sentences are randomly selected to serve as a *common* data set: they are added to each of the 9 subsets in order to observe agreements between the 9 annotators. In short, each human evaluator was presented with 42 evaluation samples. Each sample is a tuple containing the output of the baseline system, that of the system considering sentiment, and the reference translation. The two automatic translations were presented in a random order to the evaluators.

As in (Callison-Burch *et al*., 2012), we performed a pairwise comparison of the translations produced by the systems. We asked the annotators the following two questions Q1 and Q2:

- Q1(general preference): For any reason, which of the two translations do you prefer according to the provided references, otherwise mark "no preference"?
- Q2 (sentiment preference): Does the reference contains sentiment? If so, in terms of the translations of the sentiment, which of the two translations do you prefer, otherwise mark "no preference"?

We computed Fleiss's Kappa (Fleiss, 1971) on the common set to measure inter-annotator agreement, κ_{all}. Then, we excluded one and only one annotator at a time to compute κ^i (Kappa score without *i*-th annotator, *i.e.*, from the other eight). Finally, we removed the annotation of the two annotators whose answers were most different from the others': *i.e.*, annotators with the biggest $\kappa_{all} - \kappa^i$ values. As a result, we got a Kappa score 0.432 on question Q1 and 0.415 on question Q2, which both mean moderate agreement.

	base win	bsc win	equal	total
Translation	58 (31.86%)	82 (45.05%)	42 (23.09%)	182
Sentiment	30 (22.39%)	49 (36.57%)	55 (41.04%)	134

Table 7: Human evaluation preference for outputs from baseline *vs*. system with sentiment features.

This left 7 files from 7 evaluators. We threw away the common set in each file, leaving 182 pairwise comparisons. Table 6 shows that the evaluators preferred the output from the system with sentiment features 82 times, the output from the baseline system 58 times, and had no preference the other 42 times. This indicates that there is a human preference for the output from the system that incorporated the sentiment features over those from the baseline system at the $p<0.05$ significance level (in cases where people prefer one of them). For question Q2, the human annotators regarded 48 sentences as conveying no sentiment according to the provided reference, although each of them contains at least one sentiment word (a criterion we described above in constructing the evaluation set). Among the remaining 134 sentences, the human annotators preferred the proposed system 49 times and the baseline system 30 times, while they mark *no-preference* 55 times. The result shows a human preference for the proposed model that considers sentiment features at the $p<0.05$ significance level (in the cases where the evaluators did mark a preference).

4.3 Examples

We have also manually examined the translations generated by our best model (the last model of Table 4, named BSC below) and the baseline model (BSL), and we attribute the improvement to two main reasons: (1) checking sentiment consistency on a phrase pair helps punish low-quality phrase pairs caused by word alignment error, (2) such consistency checking also improves the sentiment of the translation to better match the sentiment of the source.

(1)	Phr. pairs	和谈 ||| *talks* vs. 和谈 ||| *peace talks*
	REF	… help the palestinians and the israelis to resume *peace talks* …
	BSL	… help the israelis and palestinians to resumption of the *talks* …
	BSC	… help the israelis and palestinians to resume *peace talks* …
(2)	Phr. pairs	备战 ||| *war* vs. 备战 ||| *preparing for*
	REF	… the national team is *preparing for* matches with palestine and Iraq …
	BSL	… the national team 's match with the palestinians and the iraq *war* …
	BSC	… the national team *preparing for* the match with the palestinian and iraq …
(3)	REF	… in china we have *top-quality people* , *ever-improving* facilities …
	BSL	… we have *talents* in china , an *increasing number of facilities* …
	BSC	… we have *outstanding talent* in china , *more and better* facilities …
(4)	REF	… continue to *strive* for that …
	BSL	… continue to *struggle* …
	BSC	… continue to *work hard to achieve* …

Table 8: Examples that show how sentiment helps improve our baseline model. REF is a reference translation, BSL stands for baseline model, and BSC (bilingual sentiment consistency) is the last model of Table 4.

In the first two examples of Table 8, the first line shows two phrase pairs that are finally chosen by the baseline and BSC system, respectively. The next three lines correspond to a reference (REF), translation from BSL, and that from the BSC system. The correct translations of "和谈" should be "peace negotiations" or "peace talks", which have a positive sentiment, while the word "talks" doesn't convey sentiment at all. By punishing the phrase pair "和谈 ||| talks", the BSC model was able to generate a better translation. In the second example, the correct translation of "备战" should be "prepare for", where neither side conveys sentiment. The incorrect phrase pair "备战 ||| war" is generated from incorrect word alignment. Since "war" is a negative word in our sentiment lexicon, checking sentiment consistency helps down-weight such incorrect translations. Note also that the incorrect phrase pair "备战 ||| war" is not totally irrational, as the literal translation of "备战" is "prepare for war".

Similarly, in the third example, "outstanding talent" is closer with respect to sentiment to the reference "top-quality people" than "talent" is; "more and better" is closer with respect to sentiment to the reference "ever-improving" than "an increasing number" is. These three examples also help us understand the benefit of the subjectivity features discussed in Section 3.4. In the fourth example, "work hard to achieve" has a positive sentiment, same as "strive", while "struggle" is negative. We can see that the BSC model is able to preserve the original sentiment better (the 9 human evaluators who were involved in our human evaluation (Section 4.3) all agreed with this).

5 Conclusions and future work

We explore lexicon-based sentiment consistency for statistical machine translation. By incorporating lexicon-based subjectivity, polarity, intensity, and negation features into the phrase-pair translation model, we observed a 1.1-point improvement of BLEU score on NIST Chinese-to-English translation. Among the four individual groups of features, subjectivity consistency yields the largest improvement. The usefulness of the sentiment features has also been confirmed when they are used for re-ranking, for which we observed a 0.4-point improvement on the BLEU score. In addition, human evaluation shows the preference of the human subjects towards the translations generated by the proposed model, in terms of both the general translation quality and the sentiment conveyed.

In the paper, we propose a lexicon-based approach to the problem. It is possible to employ more complicated models. For example, with the involvement of proper sentiment-annotated data, if available, one may train a better sentiment-analysis model even for the often-ungrammatical phrase pairs or sentence candidates. Another direction we feel interesting is ensuring that sentiment and its target are not only better translated but also better paired, i.e., their semantic relation is preserved. This is likely to need further syntactic or semantic analysis at the sentence level, and the semantic role labeling work reviewed in Section 2 is relevant.

References

C. Banea, R. Mihalcea, J. Wiebe and S. Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proc. of EMNLP.

E. M. Benzinger. 1971. Intensifiers in current English. PhD. Thesis. University of Florida.

P. F. Brown, S. Della Pietra, V. Della J. Pietra, and R. Mercer. 1993. The mathematics of Machine Translation: Parameter estimation. Computational Linguistics, 19(2): 263-312.

C. Callison-Burch, P. Koehn, C. Monz, R. Soricut, and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In Proc. of WMT.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proc. of ACL, 263–270.

B. Chen, G. Foster, and R. Kuhn. 2010. Bilingual Sense Similarity for Statistical Machine Translation. In Proc. of ACL, 834-843.

B. Chen, R. Kuhn, G. Foster, and H. Johnson. 2011. Unpacking and transforming feature functions: New ways to smooth phrase tables. In Proc. of MT Summit.

C. Cherry and G. Foster. 2012. Batch tuning strategies for statistical machine translation. In Proc. of NAACL.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5): 378–382.

M. Galley and C. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In Proc. of EMNLP: 848–856.

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In Proc. of EACL: 174-181.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proc. of EMNLP: 388–395.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proc. of ACL, 177-180.

J. Li, P. Resnik and H. Daume III. 2013. Modeling Syntactic and Semantic Structures in Hierarchical Phrase-based Translation. In Proc. of NAACL, 540-549.

D. Liu and D. Gildea. 2010. Semantic role features for machine translation. In Proc. of COLING, 716–724.

R. Mihalcea, C. Banea and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In Proc. of ACL.

F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In Proc. of ACL.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In Proc. of ACL.

C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. 1957. The measurement of meaning. University of Illinois Press.

B. Pang, L. Lee, S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In Proc. of EMNLP, 79-86.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of ma-chine translation. In Proc. of ACL, 311–318.

M. Taboada, M. Tofiloski, J. Brooke, K. Voll, and M. Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics. 37(2): 267-307.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proc. of HLT-NAACL, 252-259.

P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proc. of ACL, 417-424.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM based word alignment in statistical translation. In Proc. of COLING.

X. Wan. 2009. Co-Training for Cross-Lingual Sentiment Classification. In proc. of ACL, 235-243.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proc. of EMNLP.

D. Wu and P. Fung. 2009. Semantic Roles for SMT: A Hybrid Two-Pass Model. In Proc. of NAACL, 13-16.

L. Xu and H. Lin. 2007. Ontology-Driven Affective Chinese Text Analysis and Evaluation Method. In Lecture Notes in Computer Science Vol. 4738, 723-724, Springer.

C. Zhang, P. Liu, Z. Zhu, and M. Fang. 2012. A Sentiment Analysis Method Based on a Polarity Lexicon. Journal of Shangdong University (Natural Science). 47(3): 47-50.