Two Easy Improvements to Lexical Weighting

David Chiang and **Steve DeNeefe** and **Michael Pust**
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{chiang,sdeneefe,pust}@isi.edu

Abstract

We introduce two simple improvements to the lexical weighting features of Koehn, Och, and Marcu (2003) for machine translation: one which smooths the probability of translating word f to word e by simplifying English morphology, and one which conditions it on the kind of training data that f and e co-occurred in. These new variations lead to improvements of up to +0.8 BLEU, with an average improvement of +0.6 BLEU across two language pairs, two genres, and two translation systems.

1 Introduction

Lexical weighting features (Koehn et al., 2003) estimate the probability of a phrase pair or translation rule word-by-word. In this paper, we introduce two simple improvements to these features: one which smooths the probability of translating word f to word e using English morphology, and one which conditions it on the kind of training data that f and e co-occurred in. These new variations lead to improvements of up to +0.8 BLEU, with an average improvement of +0.6 BLEU across two language pairs, two genres, and two translation systems.

2 Background

Since there are slight variations in how the lexical weighting features are computed, we begin by defining the baseline lexical weighting features. If $\mathbf{f} = f_1 \cdots f_n$ and $\mathbf{e} = e_1 \cdots e_m$ are a training sentence pair, let a_i ($1 \le i \le n$) be the (possibly empty) set of positions in $\mathbf{f}$ that e_i is aligned to.

First, compute a word translation table from the word-aligned parallel text: for each sentence pair and each i, let

$$c(f_j, e_i) \leftarrow c(f_j, e_i) + \frac{1}{|a_i|} \qquad \text{for } j \in a_i \tag{1}$$

$$c(\text{NULL}, e_i) \leftarrow c(\text{NULL}, e_i) + 1 \qquad \text{if } |a_i| = 0 \tag{2}$$

Then

$$t(e \mid f) = \frac{c(f, e)}{\sum_e c(f, e)} \tag{3}$$

where f can be NULL.

Second, during phrase-pair extraction, store with each phrase pair the alignments between the words in the phrase pair. If it is observed with more than one word alignment pattern, store the most frequent pattern.

Third, for each phrase pair $(\bar{f}, \bar{e}, a)$, compute

$$t(\bar{e} \mid \bar{f}) = \prod_{i=1}^{|\bar{e}|} \begin{cases} \frac{1}{|a_i|} \sum_{j \in a_i} t(\bar{e}_i \mid \bar{f}_j) & \text{if } |a_i| > 0 \\ t(\bar{e}_i \mid \text{NULL}) & \text{otherwise} \end{cases} \tag{4}$$

This generalizes to synchronous CFG rules in the obvious way.

Similarly, compute the reverse probability $t(\bar{f} \mid \bar{e})$. Then add two new model features

$$-\log t(\bar{e} \mid \bar{f}) \quad \text{and} \quad -\log t(\bar{f} \mid \bar{e})$$

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 455–460,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

	translation	
feature	(7)	(8)
small LM	26.7	24.3
large LM	31.4	28.2
$-\log t(\bar{e} \mid \bar{f})$	9.3	9.9
$-\log t(\bar{f} \mid \bar{e})$	5.8	6.3

Table 1: Although the language models prefer translation (8), which translates 朋友 and 伙伴 as singular nouns, the lexical weighting features prefer translation (7), which incorrectly generates plural nouns. All features are negative log-probabilities, so lower numbers indicate preference.

3 Morphological smoothing

Consider the following example Chinese sentence:

(5)

温家宝	表示	，	科特迪瓦	是
Wēn Jiābǎo	biǎoshì	,	Kētèdíwǎ	shì
Wen Jiabao	said	,	Côte d'Ivoire	is

中国	在	非洲	的	好	朋友	，
Zhōngguó	zài	Fēizhōu	de	hǎo	péngyǒu	,
China	in	Africa	's	good	friend	,

好	伙伴	。
hǎo	huǒbàn	.
good	partner	.

(6) *Human:* Wen Jiabao said that Côte d'Ivoire is a good friend and a good partner of China's in Africa.

(7) *MT (baseline):* Wen Jiabao said that Cote d'Ivoire is China's good friends, and good partners in Africa.

(8) *MT (better):* Wen Jiabao said that Cote d'Ivoire is China's good friend and good partner in Africa.

The baseline machine translation (7) incorrectly generates plural nouns. Even though the language models (LMs) prefer singular nouns, the lexical weighting features prefer plural nouns (Table 1).[1]

The reason for this is that the Chinese words do not have any marking for number. Therefore the information needed to mark *friend* and *partner* for number must come from the context. The LMs are able to capture this context: the 5-gram *is China's good friend* is observed in our large LM, and the 4-gram *China's good friend* in our small LM, but *China's good friends* is not observed in either LM. Likewise, the 5-grams *good friend and good partner* and *good friends and good partners* are both observed in our LMs, but neither *good friend and good partners* nor *good friends and good partner* is.

f	e	$t(e \mid f)$	$t(f \mid e)$	$t_m(e \mid f)$	$t_m(f \mid e)$
朋友	friends	0.44	0.44	0.47	0.48
朋友	friend	0.21	0.58	0.19	0.48
伙伴	partners	0.44	0.60	0.40	0.53
伙伴	partner	0.13	0.40	0.17	0.53

Table 2: The morphologically-smoothed lexical weighting features weaken the preference for singular or plural translations, with the exception of t(friends | 朋友).

By contrast, the lexical weighting tables (Table 2, columns 3–4), which ignore context, have a strong preference for plural translations, except in the case of t(朋友 | friend). Therefore we hypothesize that, for Chinese-English translation, we should weaken the lexical weighting features' morphological preferences so that more contextual features can do their work.

Running a morphological stemmer (Porter, 1980) on the English side of the parallel data gives a three-way parallel text: for each sentence, we have French $\mathbf{f}$, English $\mathbf{e}$, and stemmed English $\mathbf{e}'$. We can then build two word translation tables, $t(e' \mid f)$ and $t(e \mid e')$, and form their product

$$t_m(e \mid f) = \sum_{e'} t(e' \mid f) t(e \mid e') \qquad (9)$$

Similarly, we can compute $t_m(f \mid e)$ in the opposite direction.[2] (See Table 2, columns 5–6.) These tables can then be extended to phrase pairs or synchronous CFG rules as before and added as two new features of the model:

$$-\log t_m(\bar{e} \mid \bar{f}) \quad \text{and} \quad -\log t_m(\bar{f} \mid \bar{e})$$

The feature $t_m(\bar{e} \mid \bar{f})$ does still prefer certain word-forms, as can be seen in Table 2. But because e is generated from e' and not from f, we are protected from the situation where a rare f leads to poor estimates for the e.

[1] The presence of an extra comma in translation (7) affects the LM scores only slightly; removing the comma would make them 26.4 and 32.0.

[2] Since the Porter stemmer is deterministic, we always have $t(e' \mid e) = 1.0$, so that $t_m(f \mid e) = t(f \mid e')$, as seen in the last column of Table 2.

When we applied an analogous approach to Arabic-English translation, stemming both Arabic and English, we generated very large lexicon tables, but saw no statistically significant change in BLEU. Perhaps this is not surprising, because in Arabic-English translation (unlike Chinese-English translation), the source language is morphologically richer than the target language. So we may benefit from features that preserve this information, while smoothing over morphological differences blurs important distinctions.

4 Conditioning on provenance

Typical machine translation systems are trained on a fixed set of training data ranging over a variety of genres, and if the genre of an input sentence is known in advance, it is usually advantageous to use model parameters tuned for that genre.

Consider the following Arabic sentence, from a weblog (words written left-to-right):

(10)	ولعل	هذا	احد	اهم	الفروق	بين
	wlEl	h*A	AHd	Ahm	Alfrwq	byn
	perhaps	this	one	main	differences	between
	صور	انظمة	الحكم	المقترحة	.	
	Swr	AnZmp	AlHkm	AlmqtrHp	.	
	images	systems	ruling	proposed	.	

(11) *Human:* Perhaps this is one of the most important differences between the images of the proposed ruling systems.

(12) *MT (baseline):* This may be one of the most important differences between pictures of the proposed ruling regimes.

(13) *MT (better):* Perhaps this is one of the most important differences between the images of the proposed regimes.

The Arabic word ولعل can be translated as *may* or *perhaps* (among others), with the latter more common according to $t(e \mid f)$, as shown in Table 3. But some genres favor *perhaps* more or less strongly. Thus, both translations (12) and (13) are good, but the latter uses a slightly more informal register appropriate to the genre.

Following Matsoukas et al. (2009), we assign each training sentence pair a set of binary features which we call *s-features*:

		$t(e \mid f)$	$t_s(e \mid f)$			
f	e	–	nw	web	bn	un
ولعل	may	0.13	0.12	0.16	0.09	0.13
ولعل	perhaps	0.20	0.23	0.32	0.42	0.19

Table 3: Different genres have different preferences for word translations. Key: nw = newswire, web = Web, bn = broadcast news, un = United Nations proceedings.

- Whether the sentence pair came from a particular genre, for example, newswire or web
- Whether the sentence pair came from a particular collection, for example, FBIS or UN

Matsoukas et al. (2009) use these s-features to compute weights for each training sentence pair, which are in turn used for computing various model features. They found that the sentence-level weights were most helpful for computing the lexical weighting features (p.c.). The mapping from s-features to sentence weights was chosen to optimize expected TER on held-out data. A drawback of this method is that we must now learn the mapping from s-features to sentence-weights and then the model feature weights. Therefore, we tried an alternative that incorporates s-features into the model itself.

For each s-feature s, we compute new word translation tables $t_s(e \mid f)$ and $t_s(f \mid e)$ estimated from only those sentence pairsf on which s fires, and extend them to phrases/rules as before. The idea is to use these probabilities as new features in the model. However, two challenges arise: first, many word pairs are unseen for a given s, resulting in zero or undefined probabilities; second, this adds many new features for each rule, which requires a lot of space.

To address the problem of unseen word pairs, we use Witten-Bell smoothing (Witten and Bell, 1991):

$$\hat{t}_s(e \mid f) = \lambda_{fs} t_s(e \mid f) + (1 - \lambda_{fs}) t(e \mid f) \qquad (14)$$

$$\lambda_{fs} = \frac{c(f, s)}{c(f, s) + d(f, s)} \qquad (15)$$

where $c(f, s)$ is the number of times f has been observed in sentences with s-feature s, and $d(f, s)$ is the number of e types observed aligned to f in sentences with s-feature s.

For each s-feature s, we add two model features

$$-\log \frac{\hat{t}_s(\bar{e} \mid \bar{f})}{t(\bar{e} \mid \bar{f})} \quad \text{and} \quad -\log \frac{\hat{t}_s(\bar{f} \mid \bar{e})}{t(\bar{f} \mid \bar{e})}$$

		Arabic-English				Chinese-English			
		newswire		web		newswire		web	
system	features	Dev	Test	Dev	Test	Dev	Test	Dev	Test
string-to-string	baseline	47.1	43.8	37.1	38.4	28.7	26.0	23.2	25.9
	full[2]	47.7	44.2*	37.4	39.0	29.5	26.8	23.8	26.3
string-to-tree	baseline	47.3	43.6	37.7	39.6	29.2	26.4	23.0	26.0
	full	47.7	44.3	38.3	40.2	29.8	27.1	23.4	26.6

Table 4: Our variations on lexical weighting improve translation quality significantly across 16 different test conditions. All improvements are significant at the $p < 0.01$ level, except where marked with an asterisk (*), indicating $p < 0.05$.

In order to address the space problem, we use the following heuristic: for any given rule, if the absolute value of one of these features is less than $\log 2$, we discard it for that rule.

5 Experiments

Setup We tested these features on two machine translation systems: a hierarchical phrase-based (string-to-string) system (Chiang, 2005) and a syntax-based (string-to-tree) system (Galley et al., 2004; Galley et al., 2006). For Arabic-English translation, both systems were trained on 190+220 million words of parallel data; for Chinese-English, the string-to-string system was trained on 240+260 million words of parallel data, and the string-to-tree system, 58+65 million words. Both used two language models, one trained on the combined English sides of the Arabic-English and Chinese-English data, and one trained on 4 billion words of English data.

The baseline string-to-string system already incorporates some simple provenance features: for each s-feature s, there is a feature $P(s \mid \text{rule})$. Both baseline also include a variety of other features (Chiang et al., 2008; Chiang et al., 2009; Chiang, 2010).

Both systems were trained using MIRA (Crammer et al., 2006; Watanabe et al., 2007; Chiang et al., 2008) on a held-out set, then tested on two more sets (Dev and Test) disjoint from the data used for rule extraction and for MIRA training. These datasets have roughly 1000–3000 sentences (30,000–70,000 words) and are drawn from test sets from the NIST MT evaluation and development sets from the GALE program.

Individual tests We first tested morphological smoothing using the string-to-string system on Chinese-English translation. The morphologically smoothed system generated the improved translation (8) above, and generally gave a small improvement:

task	features	Dev
Chi-Eng nw	baseline	28.7
	morph	29.1

We then tested the provenance-conditioned features on both Arabic-English and Chinese-English, again using the string-to-string system:

task	features	Dev
Ara-Eng nw	baseline	47.1
	(Matsoukas et al., 2009)	47.3
	provenance[2]	47.7
Chi-Eng nw	baseline	28.7
	provenance[2]	29.4

The translations (12) and (13) come from the Arabic-English *baseline* and *provenance* systems. For Arabic-English, we also compared against lexical weighting features that use sentence weights kindly provided to us by Matsoukas et al. Our features performed better, although it should be noted that those sentence weights had been optimized for a different translation model.

Combined tests Finally, we tested the features across a wider range of tasks. For Chinese-English translation, we combined the morphologically-smoothed and provenance-conditioned lexical weighting features; for Arabic-English, we continued to use only the provenance-conditioned features. We tested using both systems, and on both newswire and web genres. The results are shown in Table 4. The features produce statistically significant improvements across all 16 conditions.

[2] In these systems, an error crippled the $t(f \mid e)$, $t_m(f \mid e)$, and $t_s(f \mid e)$ features. Time did not permit rerunning all of these systems with the error fixed, but partial results suggest that it did not have a significant impact.

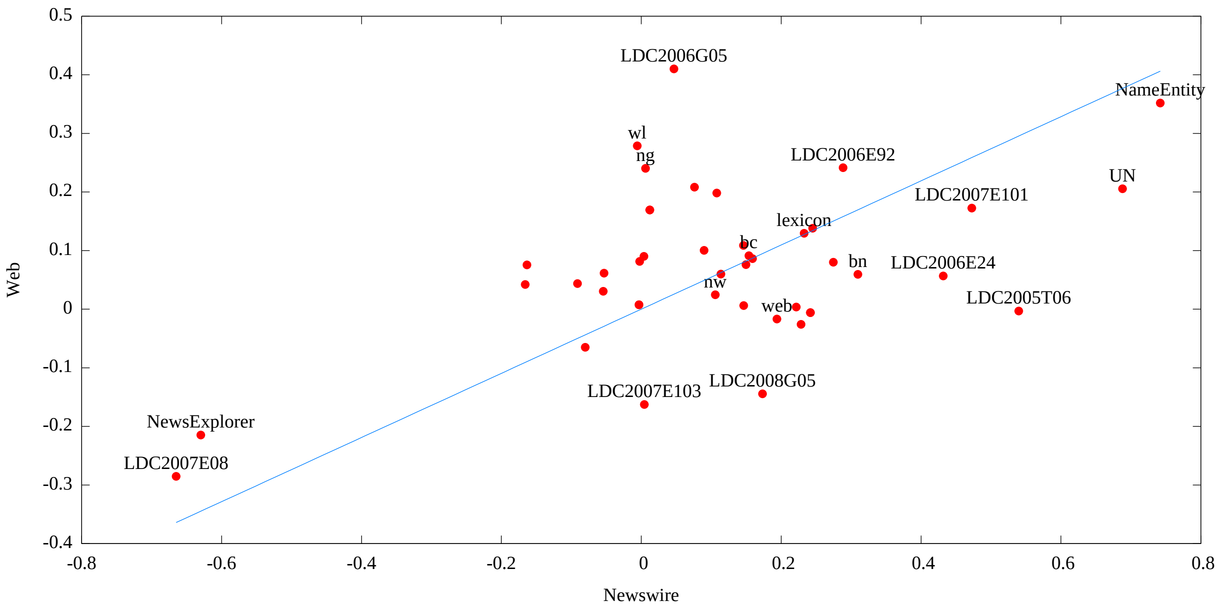

Figure 1: Feature weights for provenance-conditioned features: string-to-string, Chinese-English, web versus newswire. A higher weight indicates a more useful source of information, while a negative weight indicates a less useful or possibly problematic source. For clarity, only selected points are labeled. The diagonal line indicates where the two weights would be equal relative to the original $t(e \mid f)$ feature weight.

Figure 1 shows the feature weights obtained for the provenance-conditioned features $t_s(f \mid e)$ in the string-to-string Chinese-English system, trained on newswire and web data. On the diagonal are corpora that were equally useful in either genre. Surprisingly, the UN data received strong positive weights, indicating usefulness in both genres. Two lists of named entities received large weights: the LDC list (LDC2005T34) in the positive direction and the NewsExplorer list in the negative direction, suggesting that there are noisy entries in the latter. The corpus LDC2007E08, which contains parallel data mined from comparable corpora (Munteanu and Marcu, 2005), received strong negative weights.

Off the diagonal are corpora favored in only one genre or the other: above, we see that the `wl` (weblog) and `ng` (newsgroup) genres are more helpful for web translation, as expected (although `web` oddly seems less helpful), as well as LDC2006G05 (LDC/FBIS/NVTC Parallel Text V2.0). Below are corpora more helpful for newswire translation, like LDC2005T06 (Chinese News Translation Text Part 1).

6 Conclusion

Many different approaches to morphology and provenance in machine translation are possible. We have chosen to implement our approach as extensions to lexical weighting (Koehn et al., 2003), which is nearly ubiquitous, because it is defined at the level of word alignments. For this reason, the features we have introduced should be easily applicable to a wide range of phrase-based, hierarchical phrase-based, and syntax-based systems. While the improvements obtained using them are not enormous, we have demonstrated that they help significantly across many different conditions, and over very strong baselines. We therefore fully expect that these new features would yield similar improvements in other systems as well.

Acknowledgements

We would like to thank Spyros Matsoukas and colleagues at BBN for providing their sentence-level weights and important insights into their corpus-weighting work. This work was supported in part by DARPA contract HR0011-06-C-0022 under subcontract to BBN Technologies.

References

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. EMNLP 2008*, pages 224–233.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proc. NAACL HLT*, pages 218–226.

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

David Chiang. 2010. Learning to translate with source and target syntax. In *Proc. ACL*, pages 1443–1452.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. HLT-NAACL 2004*, pages 273–280.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. COLING-ACL 2006*, pages 961–968.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL 2003*, pages 127–133.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proc. EMNLP 2009*, pages 708–717.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Taro Watanabe, Jun Suzuki, Hajime Tsukuda, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proc. EMNLP-CoNLL 2007*, pages 764–773.

Ian H. Witten and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Information Theory*, 37(4):1085–1094.