Fixed Length Word Suffix for Factored Statistical Machine Translation

Narges Sharif Razavian
School of Computer Science
Carnegie Mellon Universiy
Pittsburgh, USA
nsharifr@cs.cmu.edu

Stephan Vogel
School of Computer Science
Carnegie Mellon Universiy
Pittsburgh, USA
stephan.vogel@cs.cmu.edu

Abstract

Factored Statistical Machine Translation extends the Phrase Based SMT model by allowing each word to be a vector of factors. Experiments have shown effectiveness of many factors, including the Part of Speech tags in improving the grammaticality of the output. However, high quality part of speech taggers are not available in open domain for many languages. In this paper we used fixed length word suffix as a new factor in the Factored SMT, and were able to achieve significant improvements in three set of experiments: large NIST Arabic to English system, medium WMT Spanish to English system, and small TRANSTAC English to Iraqi system.

1 Introduction

Statistical Machine Translation(SMT) is currently the state of the art solution to the machine translation. Phrase based SMT is also among the top performing approaches available as of today. This approach is a purely lexical approach, using surface forms of the words in the parallel corpus to generate the translations and estimate probabilities. It is possible to incorporate syntactical information into this framework through different ways. Source side syntax based re-ordering as preprocessing step, dependency based reordering models, cohesive decoding features are among many available successful attempts for the integration of syntax into the translation model. Factored translation modeling is another way to achieve this goal. These models allow each word to be represented as a vector of factors rather than a single surface form. Factors can represent richer expression power on each word. Any factors such as word stems, gender, part of speech, tense, etc. can be easily used in this framework.

Previous work in factored translation modeling have reported consistent improvements from Part of Speech(POS) tags, morphology, gender, and case factors (Koehn et. a. 2007). In another work, Birch et. al. 2007 have achieved improvement using Combinational Categorial Grammar (CCG) super-tag factors. Creating the factors is done as a preprocessing step, and so far, most of the experiments have assumed existence of external tools for the creation of these factors (i. e. Part of speech taggers, CCG parsers, etc.). Unfortunately high quality language processing tools, especially for the open domain, are not available for most languages.

While linguistically identifiable representations (i.e. POS tags, CCG supertags, etc) have been very frequently used as factors in many applications including MT, simpler representations have also been effective in achieving the same result in other application areas. Grzymala-Busse and Old 1997, DINCER et.al. 2008, were able to use fixed length suffixes as features for training a POS tagging. In another work Saberi and Perrot 1999 showed that reversing middle chunks of the words while keeping the first and last part intact, does not decrease listeners' recognition ability. This result is very relevant to Machine Translation, suggesting that inaccurate context which is usually modeled with n-gram language models, can still be as effective as accurate surface forms. Another research (Rawlinson 1997) confirms this finding; this time in textual domain, observing that randomization of letters in the middle of words has little or no effect on the ability of skilled readers to understand the text. These results suggest that the inexpensive representational factors which do not need unavailable tools might also be worth investigating.

These results encouraged us to introduce language independent simple factors for machine translation. In this paper, following the work of Grzymala-Busse et. al. we used fixed length suf-

Proceedings of the ACL 2010 Conference Short Papers, pages 147–150,
Uppsala, Sweden, 11-16 July 2010.

fix as word factor, to lower the perplexity of the language model, and have the factors roughly function as part of speech tags, thus increasing the grammaticality of the translation results. We were able to obtain consistent, significant improvements over our baseline in 3 different experiments, large NIST Arabic to English system, medium WMT Spanish to English system, and small TRANSTAC English to Iraqi system.

The rest of this paper is as follows. Section 2 briefly reviews the Factored Translation Models. In section 3 we will introduce our model, and section 4 will contain the experiments and the analysis of the results, and finally, we will conclude this paper in section 5.

2 Factored Translation Model

Statistical Machine Translation uses the log linear combination of a number of features, to compute the highest probable hypothesis as the translation.

$$e = \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}}\ p\ exp\ \Sigma_{i=1}^{n}\ \lambda_i\ h_i(\boldsymbol{e},\boldsymbol{f})$$

In phrase based SMT, assuming the source and target phrase segmentation as $\{(f_i,e_i)\}$, the most important features include: the Language Model feature $h_{lm}(\boldsymbol{e},\boldsymbol{f}) = p_{lm}(\boldsymbol{e})$; the phrase translation feature $h_t(\mathbf{e},\mathbf{f})$ defined as product of translation probabilities, lexical probabilities and phrase penalty; and the reordering probability, $h_d(\mathbf{e},\mathbf{f})$, usually defined as $\pi_{i=1}^{n}\ d(start_i,end_{i-1})$ over the source phrase reordering events.

Factored Translation Model, recently introduced by (Koehn et. al. 2007), allow words to have a vector representation. The model can then extend the definition of each of the features from a uni-dimensional value to an arbitrary joint and conditional combination of features. Phrase based SMT is in fact a special case of Factored SMT.

The factored features are defined as an extension of phrase translation features. The function $\tau(f_j,e_j)$, which was defined for a phrase pair before, can now be extended as a log linear combination $\Sigma_f\ \tau_f(\underline{f}_{jf},\underline{e}_{jf})$. The model also allows for a generation feature, defining the relationship between final surface form and target factors. Other features include additional language model features over individual factors, and factored reordering features.

Figure 1 shows an example of a possible factored model.

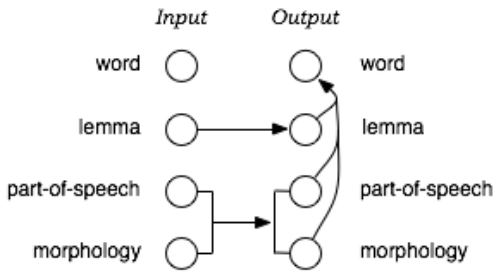

Figure 1: An example of a Factored Translation and Generation Model

In this particular model, words on both source and target side are represented as a vector of four factors: surface form, lemma, part of speech (POS) and the morphology. The target phrase is generated as follows: Source word lemma generates target word lemma. Source word's Part of speech and morphology together generate the target word's part of speech and morphology, and from its lemma, part of speech and morphology the surface form of the target word is finally generated. This model has been able to result in higher translation BLEU score as well as grammatical coherency for English to German, English to Spanish, English to Czech, English to Chinese, Chinese to English and German to English.

3 Fixed Length Suffix Factors for Factored Translation Modeling

Part of speech tagging, constituent and dependency parsing, combinatory categorical grammar super tagging are used extensively in most applications when syntactic representations are needed. However training these tools require medium size treebanks and tagged data, which for most languages will not be available for a while. On the other hand, many simple words features, such as their character n-grams, have in fact proven to be comparably as effective in many applications.

(Keikha et. al. 2008) did an experiment on text classification on noisy data, and compared several word representations. They compared surface form, stemmed words, character n-grams, and semantic relationships, and found that for noisy and open domain text, character-ngrams outperform other representations when used for text classification. In another work (Dincer et al 2009) showed that using fixed length word ending outperforms whole word representation for training a part of speech tagger for Turkish language.

Based on this result, we proposed a suffix factored model for translation, which is shown in Figure 2.

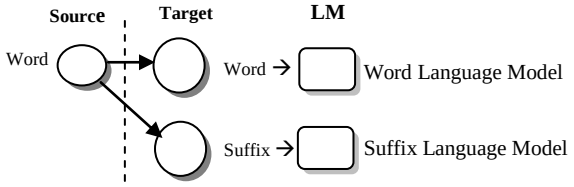

Figure 2: Suffix Factored model: Source word determines factor vectors (target word, target word suffix) and each factor will be associated with its language model.

Based on this model, the final probability of the translation hypothesis will be the log linear combination of phrase probabilities, reordering model probabilities, and each of the language models' probabilities.

$$P(e|f) \sim p_{lm\text{-}word}(e_{word}) * p_{lm\text{-}suffix}(e_{suffix}) * \Sigma_{i=1}^{n} p(e_{word\text{-}j} \,\&\, e_{suffix\text{-}j}|f_j) * \Sigma_{i=1}^{n} p(f_j \mid e_{word\text{-}j} \,\&\, e_{suffix\text{-}j})$$

Where $p_{lm\text{-}word}$ is the n-gram language model probability over the word surface sequence, with the language model built from the surface forms. Similarly, $p_{lm\text{-}suffix}(e_{suffix})$ is the language model probability over suffix sequences. $p(e_{word\text{-}j} \,\&\, e_{suffix\text{-}j}|f_j)$ and $p(f_j \mid e_{word\text{-}j} \,\&\, e_{suffix\text{-}j})$ are translation probabilities for each phrase pair i , used in by the decoder. This probability is estimated after the phrase extraction step which is based on grow-diag heuristic at this stage.

4 Experiments and Results

We used Moses implementation of the factored model for training the feature weights, and SRI toolkit for building n-gram language models. The baseline for all systems included the moses system with lexicalized re-ordering, SRI 5-gram language models.

4.1 Small System from Dialog Domain: English to Iraqi

This system was TRANSTAC system, which was built on about 650K sentence pairs with the average sentence length of 5.9 words. After choosing length 3 for suffixes, we built a new parallel corpus, and SRI 5-gram language models for each factor. Vocabulary size for the surface form was 110K whereas the word suffixes had about 8K distinct words. Table 1 shows the result (BLEU Score) of the system compared to the baseline.

System	Tune on Set-July07	Test on Set-June08	Test on Set-Nov08
Baseline	27.74	21.73	15.62
Factored	28.83	22.84	16.41
Improvement	1.09	1.11	0.79

Table 1: BLEU score, English to Iraqi Transtac system, comparing Factored and Baseline systems.

As you can see, this improvement is consistent over multiple unseen datasets. Arabic cases and numbers show up as the word suffix. Also, verb numbers usually appear partly as word suffix and in some cases as word prefix. Defining a language model over the word endings increases the probability of sequences which have this case and number agreement, favoring correct agreements over the incorrect ones.

4.2 Medium System on Travel Domain: Spanish to English

This system is the WMT08 system, on a corpus of 1.2 million sentence pairs with average sentence length 27.9 words. Like the previous experiment, we defined the 3 character suffix of the words as the second factor, and built the language model and reordering model on the joint event of (surface, suffix) pairs. We built 5-gram language models for each factor. The system had about 97K distinct vocabulary in the surface language model, which was reduced to 8K using the suffix corpus. Having defined the baseline, the system results are as follows.

System	Tune-WMT06	Test set-WMT08
Baseline	33.34	32.53
Factored	33.60	32.84
Improvement	0.26	0.32

Table 2: BLEU score, Spanish to English WMT system, comparing Factored and Baseline systems.

Here, we see improvement with the suffix factors compared to the baseline system. Word endings in English language are major indicators of word's part of speech in the sentence. In fact

most common stemming algorithm, Porter's Stemmer, works by removing word's suffix. Having a language model on these suffixes pushes the common patterns of these suffixes to the top, making the more grammatically coherent sentences to achieve a better probability.

4.3 Large NIST 2009 System: Arabic to English

We used NIST2009 system as our baseline in this experiment. The corpus had about 3.8 Million sentence pairs, with average sentence length of 33.4 words. The baseline defined the lexicalized reordering model. As before we defined 3 character long word endings, and built 5-gram SRI language models for each factor. The result of this experiment is shown in table 3.

System	Tune on MT06	Test on Dev07 News Wire	Test on Dev07 Weblog	Test on MT08
Baseline	43.06	48.87	37.84	41.70
Factored	44.20	50.39	39.93	42.74
Improvement	1.14	1.52	2.09	1.04

Table 3: BLEU score, Arabic to English NIST 2009 system, comparing Factored and Baseline systems.

This result confirms the positive effect of the suffix factors even on large systems. As mentioned before we believe that this result is due to the ability of the suffix to reduce the word into a very simple but rough grammatical representation. Defining language models for this factor forces the decoder to prefer sentences with more probable suffix sequences, which is believed to increase the grammaticality of the result. Future error analysis will show us more insight of the exact effect of this factor on the outcome.

5 Conclusion

In this paper we introduced a simple yet very effective factor: fixed length word suffix, to use in Factored Translation Models. This simple factor has been shown to be effective as a rough replacement for part of speech. We tested our factors in three experiments in a small, English to Iraqi system, a medium sized system of Spanish to English, and a large system, NIST09 Arabic to English. We observed consistent and significant improvements over the baseline. This result, obtained from the language independent and inexpensive factor, shows promising new opportunities for all language pairs.

References

Birch, A., Osborne, M., and Koehn, P. CCG supertags in factored statistical machine translation. Proceedings of the Second Workshop on Statistical Machine Translation, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics, 2007.

Dincer T., Karaoglan B. and Kisla T., A Suffix Based Part-Of-Speech Tagger For Turkish, Fifth International Conference on Information Technology: New Generations, 2008.

Grzymala-Busse J.W., Old L.J. A machine learning experiment to determine part of speech from word-endings, Lecture Notes in Computer Science, Communications Session 6B Learning and Discovery Systems, 1997.

Keikha M., Sharif Razavian N, Oroumchian F., and Seyed Razi H., Document Representation and Quality of Text: An Analysis, Chapter 12, Survey of Text Mining II, Springer London, 2008.

Koehn Ph., Hoang H., Factored Translation Models, Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), 2007.

Rawlinson G. E., The significance of letter position in word recognition, PhD Thesis, Psychology Department, University of Nottingham, Nottingham UK, 1976.

Saberi K and Perrot D R, Cognitive restoration of reversed speech, Nature (London) 1999.