Augmenting Manual Dictionaries for Statistical Machine Translation Systems

Stephan Vogel, Christian Monson
Language Technologies Institute,
Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3891 U.S.A.
vogel+,cmonson@cs.cmu.edu

Abstract

We show that the usefulness of manually created dictionaries can be enhanced for a statistical machine translation system when new translations are automatically added which are simple morphological transformations (plural forms, different verb inflections) of the original. Further improvement is possible when assigning probabilities to the lexicon entries. We describe a method to do this on the basis of an automatically trained statistical lexicon. Experimental results are given for Chinese to English translation tasks and show a significant improvement in translation quality.

Introduction

Manual dictionaries are valuable resources in automatic machine translation and they can be used to improve statistical machine translation (SMT) systems. In our Chinese-to-English translation system, employing a dictionary distributed by LDC gave a significant improvement. Error analysis of the system's output revealed, however, that very often the translations from the manual dictionary, though correct in their base form, were missing the article in the case of nouns, had the singular form where the plural was required, or had the wrong verb form. Dictionaries usually contain only entries for the base form, not for inflected word forms. The SMT system, on the other hand, works with full word forms. This led to the idea to augment the dictionary with additional word forms and to add definite and indefinite articles to noun phrases. This can be done automatically and involves only part-of-speech information on the English side, which is readily available.

A second draw-back of using a manually created dictionary is that the entries do not have information on how likely the different translation alternatives are. It is up to the language model used in the SMT system to select one of the translation alternatives. In this paper we investigate the possibilities of adding probabilities to the dictionary based on word-pair frequencies observed in a bilingual corpus.

The next section describes the augmentation of the dictionary. This is followed by a proposal for assigning probabilities to all entries in the dictionary. We then report translation results which demonstrate the effect of augmentation and adding probabilities to the dictionary.

Augmenting the Dictionary

LDC distributes a Chinese-to-English dictionary, which has 54,131 Chinese entries with a total of 81,945 Chinese-English translation pairs. In the so-called small data track evaluation in the TIDES project a subset of this dictionary is used, which has 10K Chinese lexical items and 21,486 translation pairs.

Adding new translations for the lexical entries is a two steps process:

1. Simple morphological variations are automatically generated based on word class information by:
   - Identifying the parts-of-speech for the English translation. Multiple POS tags are allowed, e.g. noun and verb;
   - For nouns and noun phrases: generating plural forms and entries with definite and indefinite determiners;
   - For verbs: generating -s -ed and -ing forms, also the infinitive form with 'to'.
2. A large monolingual English corpus is used to filter the new word forms (not entire entries): if they do not appear in the corpus, the new entries are not added to the lexicon.

To identify the POS of words on the English side of the LDC Chinese-to-English dictionary we consult a word list derived from the British National Corpus (BNC). This wordlist contains of 130,000 English words tagged with POS information. We use the basic POS tag set with 61 tags. When a word is tagged with several tags, e.g. noun and verb, or adjective and noun, additional lexicon entries for each POS are generated.

For words which are in the LDC dictionary, but not in the BNC, no additional entries are generated. For the full LDC dictionary, this is the case for about 7,000 words out of 28,000 words. Of the 9,000 words in the 10K dictionary only 835 words are not covered by the BNC word list.

Starting from the original LDC dictionary with the 81,945 Chinese-English translation pairs, adding these additional entries resulted in an augmented dictionary with 420,033 translation pairs. For the 10K dictionary the augmentation increased the number of translation pairs from 21,486 to 146,099.

It should be mentioned that augmentation can introduce unwanted translations, especially, as we do not

Table 1: Translation Probabilities for the Full LDC Dictionary

	部	部门	处	科系	系	学系
Department	0.183	0.033	0.112	6.2e-9	6.2e-9	6.2e-9
Departments	0.066	0.530	0.001	6.2e-9	1.5e-4	6.2e-9
Division	0.024	0.003	--	--	--	
Divisions	0.038	0.023	--	--	--	--
Ministry	0.421	--	--	--	--	--
Ministries	0.285	--	--	--	--	--
Faculty	--	--	--	--	0.327	--
Faculties	--	--	--	--	0.083	--
Office	--	--	0.094	--	--	--
Offices	--	--	0.099	--	--	--
School	--	--	--	--	--	0.385
Schools	--	--	--	--	--	0.608
Science	--	--	--	0.372	--	--
Sciences	--	--	--	0.273	--	--
Section	0.032	1e-4	--	--	--	--
Sections	0.051	0.031	--	--	--	--

distinguish between upper and lower case. The entry for the noun ‘March’, for example, was augmented with 14 different forms of the verb ‘march’ like ‘I march’ or ‘they marched’.

Assigning Probabilities

Augmenting the dictionary with additional translations increases the need for a good strategy of selecting an appropriate entry when translating a sentence. As we use the lexicon in the context of statistical machine translation, a language model for the target language is used to select one out of several alternatives. Here, we propose to assign probabilities to the translation pairs in the lexicon. This can be done by using co-occurrence information from bilingual corpora. Using a standard word alignment model a statistical word-to-word lexicon can be trained. The probabilities for the translation pairs in the augmented manual lexicon, which can be multi-word to multi-word translations, are then calculated according to

$$p(f \mid e) = \prod_j \sum_i p(f_j \mid e_i)$$

i.e. product over source words f_j and, for each source word, sum of the word-to-word translation probabilities p(f | e) over all target words e_i. For the LDC dictionary we typically have only one source word, but often several target words.

To calculate the probabilities for the 10K dictionary a statistical lexicon was trained on a small corpus containing only about 3,500 sentence pairs, in line with the definition of the small data track conditions for the TIDES machine translation evaluation. Of the 9987 Chinese words only 5,477 were seen in this training corpus, and of the 9,061 English words only 4,551 appeared in the training data. This indicates that for many of the translation pairs in the 10K LDC dictionary only default probabilities could be assigned. For the full dictionary a large training corpus was used. Therefore, the coverage of the entries in the manual dictionary is higher. But still, 13,913 out of 46,332 Chinese words and 10,545 out of 28,203 English words were not covered by the training corpus.

We could avoid having entries with small probabilities by adding the manual dictionary to the bilingual training corpus from which the probabilities for the statistical lexicon are estimated. However, the augmentation of the dictionary introduces some wrong translation pairs, and those would then be assigned a high probability.

The probabilities for the translation pairs can be used as given by the above equation, or they can be renormalized. When using the probabilities as given the manual lexicon is well balanced with the statistical lexicon and the phrase translation probabilities. This is usually the preferred situation. When using only the manual lexicon renormalization can give slightly better results when most of the entries get only the small default probabilities.

Example

An example will show the effect of augmentation and also the probabilities which are assign to different lexical entries.

There are 6 Chinese entries which have as one of their translations ‘department’.

部	- department, division, ministry, section
部门	- department, division, section
处	- department, offices
科系	- department, science
系	- department, faculty
学系	- department, school

Some of these words have even more translations, but to illustrate our approach, these suffice. For each translation the plural form is added, as well as translations with definite and indefinite articles for the singular form and a definite article for the plural form.

Using the statistical lexicon the probabilities assigned to the different translations are as given in Table 1. We observe the following:

1. Only for 3 out of 6 Chinese words do we get a high translation probability for 'department' or 'departments'. But the other 3 words have high probabilities with other translations.
2. The probabilities for singular forms and plural forms are usually different, where in some cases, e.g. for 部门, the plural forms have higher probabilities.
3. Some entries have a very small probability 6.2×10^{-9}. The word pairs have not been seen in the training corpus, and therefore a small default probability is assigned. This value depends on the smoothing of the statistical lexicon.

It should also be mentioned that the probabilities for the translations with articles (not shown in Table 1) differ not significantly from those without article, as the probabilities p(Chinese word | English article) are typically very small. The language model has to choose between those alternatives.

Some of the English translations in Table 1 are also treated as verbs, like 'section' and 'school'. This leads to additional entries in the augmented dictionary like:

部门 - to section, I section, I sectioned, he sections;

Again, these entries have typically the same translation probability as the ones generated from the noun and we have to rely on the language model to select the correct translation which, of course, is not guaranteed.

Experiments

To study the effect of augmenting the dictionary and assigning probabilities to the entries we ran a number of experiments on the test data used in the June-2002 TIDES machine translation evaluation. This test set consists of 100 news stories, adding up to 878 sentences. We use the NIST mteval metric to measure the translation quality (NIST Report 2002), with four reference translations.

The statistical translation system used in these experiments has been described in detail in (Vogel et al. 2003, Vogel 2003). It uses a 3-gram language model in addition to the translation model. The translation model is typically using phrase translation pairs which are extracted from a bilingual training corpus.

We report results for a small data scenario, which uses a bilingual corpus of about 100K words and the 10K LDC dictionary, and for a large data scenario, which uses a training corpus of about 100 million words and the full dictionary. The statistical lexicon is generated by applying the IBM1 word alignment model (Brown et al. 1993). Other word alignment models could be used to estimate the lexical probabilities p(f | e). The IBM1 model has the advantage that it is simple and leads to a global optimum in the Expectation-Maximization training.

In the first experiment only the LDC dictionaries were used. All runs used the same parameter settings, esp. the same scaling factor for the language model. Untranslated words were deleted from the output. Using different parameter settings results in slightly different scores, but the overall picture stays the same. Table 2 shows the NIST scores under different conditions.

Table 2: Translation Results for June 2002 Test Set

	10K	Full
orig. LDC, no LM	3.79	3.72
orig. LDC, with LM	5.40	5.52
augm. LDC, no LM	3.93	3.49
augm. LDC, with LM	5.78	6.15
augm. LDC, probs renorm, no LM	3.93	4.23
augm. LDC, probs renorm, with LM	**5.91**	6.28
augm. LDC, probs no-ren, with LM	4.77	**6.59**

Without a language model and without translation probabilities the first translation will always be picked by the decoder. Augmenting the dictionary provides some useful new translations but they are only selected appropriately when the LM is added, helping the system to discriminate between good and bad augmentations. Actually, without an LM the performance can even drop, as the first translation, which depends on the sorting of the dictionary, might in some cases be worse than the first translation in the original dictionary.

Best results were achieved when also assigning probabilities to the translation pairs. Renormalization of the probabilities gave a better result for the small dictionary, whereas the full dictionary gave the best results when using the translation probabilities as calculated on the basis of the statistical lexicon. In the case of the 10K dictionary many entries have only the very small probability resulting from the default probability of unseen word pairs. The translation system prefers to output the source word rather than an unlikely English word. As untranslated words are removed from the output the translations tend to be too short, resulting in a rather high length penalty from the NIST metric. Renormalization leads to larger probabilities, lessening this effect and leading to higher translation scores.

Overall we see an improvement of 0.38 and 0.63 in NIST score resulting from the augmentation alone. Adding probabilities to the manual dictionaries allows the translation model to be more discriminative and gives an additional improvement of 0.13 and 0.44. The overall improvement when using morphological augmentation and probabilities amounts to 0.51 in NIST score for the 10K dictionary and 1.07 for the full dictionary. All these improvements except the 0.13 are statistically significant on the 95% level, using the bootstrap technique (Zhang et al. 2004) to test significance.

In the final experiment the effect of the augmented dictionary in a full statistical translation system was studied. The full SMT system uses word-to-word and phrase-to-phrase translations, extracted automatically from the bilingual training corpus (Vogel et al. 2003, Zhang et al. 2003). Table 3 gives the results for both the small and the large data system. The LM is used in all translation runs.

Table 3: Effect of LDC Dictionary in Full Translation System

	10K	Full
Baseline	5.96	6.80
+ orig. LDC	6.41	7.08
+ augm. LDC	6.66	7.11
+ augm. LDC probs. renorm.	**6.71**	7.35
+ augm. LDC probs. no-renorm.	6.05	**7.66**

The baseline system uses only the word and phrase translations learned from the bilingual corpus. We see that adding the manual dictionary gives already some improvement. Whereas there is hardly any effect in the large data system augmentation does help in the small data system. Adding probabilities, however, leads to further improvement for both systems. Overall, the effect of the manual dictionary is less pronounced, as fewer words are translated based on the dictionary. And the improvements are more significant for the small data system, where the vocabulary coverage from the training data is smaller, and hence, more words in the test sentences are translated using the manual dictionary.

Again, we see that for the 10K dictionary renormalization of the probabilities is important. Without renormalization the translations provided by the LDC dictionary have often much smaller probabilities than those provided from the statistical lexicon and the phrase translation pairs and are therefore not selected. For the large data system with the full LDC dictionary, the probabilities without renormalization are more reliable and reasonably well balanced with the probabilities of the other translations.

Conclusion

In this paper we studied the effect of augmenting a given manual lexicon with automatically generated translations, using simple morphological variations. In addition we used co-occurrence frequencies collected from bilingual data to assign translation probabilities to the lexicon entries. Both extensions to the original lexicon resulted in significant improvements in translation quality, not only when using the dictionary alone, but also when using the dictionary in a full statistical translation system.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19, no. 2, pp. 263--311, 1993.

British National Corpus: http://www.natcorp.ox.ac.uk/

NIST Report (2002) Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, Alex Waibel (2003) The CMU Statistical Translation System. *Proceedings of MT Summit IX*, New Orleans, LA, U.S.A., September 2003.

Stephan Vogel (2003) SMT Decoder Dissected: Word Reordering. *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'03)*, Beijing, China. October 2003.

Ying Zhang, Stephan Vogel, Alex Waibel (2003) Integrated Phrase Segmentation and Alignment Algorithm for Statistical Machine Translation. *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'03)*, Beijing, China, October 2003.

Ying Zhang, Stephan Vogel, Alex Waibel (2004) Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System? *Proceedings of LREC 2004*, Lisbon, Portugal, Mai 2004