Exploiting Phrasal Lexica and Additional Morpho-syntactic Language Resources for Statistical Machine Translation with Scarce Training Data

Maja Popović and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{popovic,ney}@informatik.rwth-aachen.de

Abstract. In this work, the use of a phrasal lexicon for statistical machine translation is proposed, and the relation between data acquisition costs and translation quality for different types and sizes of language resources has been analyzed. The language pairs are Spanish-English and Catalan-English, and the translation is performed in all directions. The phrasal lexicon is used to increase as well as to replace the original training corpus. The augmentation of the phrasal lexicon with the help of additional monolingual language resources containing morpho-syntactic information has been investigated for the translation with scarce training material. Using the augmented phrasal lexicon as additional training data, a reasonable translation quality can be achieved with only 1000 sentence pairs from the desired domain.

1 Introduction and Related Work

The goal of statistical machine translation (SMT) is to translate an input word sequence $f_1^J = f_1 \dots f_j \dots f_J$ into a target word sequence $e_1^I = e_1 \dots e_i \dots e_I$ by maximising the probability $P(e_1^I|f_1^J)$. This probability can be factorised into the translation model probability $P(f_1^J|e_1^I)$, which describes the correspondence between the words in the source and the target sequence and the language model probability $P(e_1^J)$, which describes the well-formedness of the produced target sequence. These two probabilities can be modelled independently of each other. For detailed descriptions of SMT models, see for example (Brown et al., 1993). Translation probabilities are extracted from a bilingual parallel text corpus, whereas language model probabilities are learnt from a monolingual text corpus in the target language. Usually, the larger the available training corpus, the better the performance of a translation system. However, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires lot of time and effort, and, for many language pairs, is not even possible. Therefore, the strategies for exploiting limited amounts of bilingual data are receiving more and more attention (Al-Onaizan et al., 2000; Nießen and Ney, 2004; Matusov et al., 2004).

Conventional dictionaries (one word and its translation(s) per entry) have been proposed in (Brown et al., 1993) and are shown to be valuable resources for SMT systems. They can be used to augment and also to replace the training corpus. Nevertheless, the main draw-back is that they typically contain only base forms of the words and not inflections. The use of morpho-syntactic information for overcoming this problem is investigated in (Nießen and Ney, 2004) for translation from German into English and in (Vogel and Monson, 2004) for translation from Chinese into English. Still, the dictionaries normally contain one word per entry and do not take into account phrases, idioms and similar complex expressions.

In our work, we have exploited a phrasal lexicon (one short phrase and its translation(s) per entry) as a bilingual knowledge source for SMT which has not been examined so far. A phrasal lexicon is expected to be especially helpful to overcome some difficulties which cannot be handled well with standard dictionaries.

We have used the phrasal lexicon to increase the existing training corpus as well as to replace it. We have also investigated the augmentation of the lexicon by the use of additional morpho-syntactically annotated language resources in order to obtain a

reasonable translation quality with minimal amount of training data.

The language pairs in our experiments are Spanish-English and Catalan-English, and translation is performed in all four directions using the phrase-based SMT system with optimised scaling factors (Och and Ney, 2002).

2 Language Resources

The parallel trilingual corpus used in our experiments has been successively built in the framework of the LC-STAR project (Arranz et al., 2003). It consists of spontaneous dialogues in Spanish, Catalan and English in the tourism and travelling domain. The development and test set are randomly extracted from this corpus and the rest is used for training (referred to as 40k).

In order to investigate the scenario with scarce training material, a small training corpus (referred to as 1k) has been constructed by random selection of 1000 sentences from the original trilingual training set.

2.1 Phrasal Lexicon

The phrasal lexicon used in our experiments (PL) consists of a list of English phrases and their translations into Spanish and Catalan. These English phrases have been extracted partly from various dialogue corpora and web-sites and have partly been created manually. The average phrase length is short, with about 4 words per entry. However, the vocabularies are rather large for all three languages, as can be seen in the Table 1. Besides full forms of the words, POS tags for all three languages as well as base forms for Spanish and Catalan are also available.

2.2 Word Expansion Lists

For Spanish and Catalan, there was an additional monolingual language resource available: a list of word base forms along with all possible POS tags and all full forms that can be derived from them. This expansion list was extracted from the morphological analyzer by UPC (Carmona et al., 1998).

Since Spanish and Catalan have an especially rich morphology for verbs, we used these lists only for verb expansions for the experiments reported in this paper. However, it might be reasonable to include expansions of other word classes in some future experiments.

3 Experiments and Results

3.1 Experimental Settings

The experiments have been done on the full training corpus containing about 40k sentences and 500k running words as well as on the small training corpus containing about 1k sentences and 12k running words. We also present the results obtained using only the phrasal lexicon as training corpus. The corpus statistics is shown in Table 1.

Extensions of the phrasal lexicon have been done using base forms and POS information for the Spanish and Catalan verbs and word expansion lists for those two languages. Each base form of the verb seen in the lexicon more than five times has been expanded with all POS tags seen in the lexicon. For each base form and POS tag, the possible English equivalents are extracted using lexical probabilities and then manually checked and eventually corrected. For example, for the Spanish base form and POS tag combination “ir VMIP1S0”, the correct English equivalents are “I go” and “I am going”. Finally, each base form and POS tag is mapped to the corresponding full form by using the word expansion list, e.g. “ir VMIP1S0” is mapped to “voy” and “ir VMIF1P0” is mapped to “iremos” (we will go). In this way, the lexicon was enriched with some previously unseen full forms of the verb.

The translation with the system trained only on the phrasal lexicon is done without a language model as well as with the language model trained on the full target language corpus. The same set-up has been used for the small training corpus - once with the small language model trained on 1k sentences and once with the full language model trained on 40k sentences.

In order to investigate the effects of the phrasal lexicon and the size of bilingual corpus available for training on translation quality, the following set-ups have been defined:

1. full training corpus;
2. full training corpus with phrasal lexicon;
3. small training corpus;

4. small training corpus with phrasal lexicon;
5. small training corpus with extended phrasal lexicon;
6. small training corpus with extended phrasal lexicon and language model trained on the full corpus;
7. phrasal lexicon without language model;
8. extended phrasal lexicon without language model;
9. extended phrasal lexicon and full language model.

Besides the standard development and test set described in Section 2, we also performed translation on an external test text which does not come from the same domain as the training set. This text has been translated with the systems 1, 3, 6, 7 and 9.

3.2 Results

The translation results can be seen in Table 2 and Table 3. As expected, the best results are obtained using the full training corpus with additional phrasal lexicon. It can be seen that the use of the phrasal lexicon yields improvements of the translation quality even when a large bilingual corpus from the domain is available, although these improvements are relatively small.

However, for the small bilingual corpus, the importance of the phrasal lexicon is significant. The degradation in terms of WER and PER by using only 2.5% of the original corpus is not higher than 25% relative if the additional phrasal lexicon and language resources containing morphological information are available. This can be further improved by up to 1.9% absolute in WER if monolingual in-domain data is available, so that a better language model can be trained. This effect is even more significant for the external test corpus, although all the error rates are higher since this text does not come from the same domain. As can be seen in Table 4, the degradation of error rates by reducing the training corpus is not higher than 6% relative if the extended phrasal lexicon is added. The big advantage of using such a small corpus is that its acquisition should not require any particular effort since producing 1000 parallel sentences in two or three languages can also be done manually.

Using only the phrasal lexicon and additional resources, the obtained error rates are similar to those for the small training corpus alone. These error rates are rather high, but they might be acceptable for tasks where only the gist of the translated text is needed, like for example document classification or multi-lingual information retrieval.

Some translation examples for the direction English→Spanish using only the phrasal lexicon with and without verb expansions are shown in Table 5. It can be seen that the extension of the lexicon enables the system to find the correct full form of the verb in the inflected language more often. For the translation Spanish→English, Table 6 shows that, for the extended lexicon, the system is able to produce correct or approximatively correct translations even for full forms that have not been seen in the original training corpus. In the baseline system, those words remain untranslated and are marked by UNKNOWN_. The effects of the lexicon extension for the other language pair (Catalan-English) as well as for the scenarios with the small training corpus are basically the same.

4 Conclusion

In this work, we have examined the possibilities for the use of a phrasal lexicon for statistical machine translation, especially as an additional resource for scarce training material.

We presented different translation scenarios: with the full training corpus, with only 1000 sentence pairs (2.5% of the full corpus), both with and without phrasal lexicon as additional data and also only with the phrasal lexicon as training corpus. We also studied the effects of extending a lexicon using morpho-syntactic knowledge sources. We showed that with the extended phrasal lexicon as additional training data, an acceptable translation quality can be achieved with only 1000 sentence pairs of in-domain text. The big advantage of such a small corpus is that the costs of its acquisition are rather low - such a corpus basically can be produced manually in relatively short time.

In our future research, we plan to examine combinations of phrasal lexica and conventional dictio-

Table 1. Corpus Statistics

		Spanish	Catalan	English
Training:				
full corpus (40k)	Sentences	40574		
	Running Words + Punct.	482290	485514	516717
	Vocabulary	14327	12772	8116
	Singletons	6743	5930	3081
small corpus (1k)	Sentences	1014		
	Running Words + Punct.	12138	12215	12972
	Vocabulary	1880	1823	1436
	Singletons	1150	1070	744
phrasal lexicon (PL)	Entries	10520		
	Running Words + Punct.	44289	46002	41850
	Vocabulary	10797	10460	11167
	Singletons	6573	6218	7153
Development:	Sentences	972		
	Running Words + Punct.	12883	13039	13983
	OOVs - 40k	209 (1.4%)	179 (1.4%)	95 (1.2%)
	OOVs - 1k	1105 (58.8%)	1029 (56.4%)	766 (53.3%)
	OOVs - PL	726 (6.7%)	627 (6.0%)	328 (2.9%)
Test:	Sentences	972		
	Running Words + Punct.	12771	12973	13922
	OOVs - 40k	206 (1.4%)	171 (1.3%)	117 (1.4%)
	OOVs - 1k	1095 (58.2%)	1008 (55.3%)	777 (54.1%)
	OOVs - PL	733 (6.8%)	611 (5.8%)	365 (3.1%)
External Test:	Sentences	200		
	Running Words + Punct.	2949	3020	3117
	OOVs - 40k	19 (0.1%)	36 (0.3%)	59 (0.7%)
	OOVs - 1k	275 (14.6%)	284 (15.6%)	234 (16.3%)
	OOVs - PL	176 (1.6%)	184 (1.8%)	107 (0.9%)

naries, and also to investigate effects for other language pairs and tasks.

Acknowledgement

This work was partly funded by the TC-STAR project by the European Community (FP6-506738) and by the Deutsche Forschungsgemeinschaft (DFG) under the project “Statistical Methods for Written Language Translation” (Ne572/5).

5 References

Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, K. Yamada. 2000. Translating with scarce resources. In *Proc. of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 672–678, Austin, TX, August.

V. Arranz, N. Castell, and J. Giménez. 2003. Development of language resources for speech-to-speech translation. In *Proc. of RANLP’03*, Borovets, Bulgaria, September.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and M. J. Goldsmith. 1993. But Dictionaries are Data Too. In *Proc. ARPA Human Language Technology Workshop ’93*, pages 202–205, Princeton, NJ, March.

J. Carmona, S. Cervell, L. Màrquez, M. Martí,

Table 2. Translation Error Rates [%] for the language pair Spanish-English

Spanish→English	Development			Test		
Training Corpus	WER	PER	1-BLEU	WER	PER	1-BLEU
40k (full corpus)	40.8	33.3	59.9	41.5	33.9	60.8
+PL	39.5	32.1	58.9	40.8	32.8	59.9
1k (small corpus)	53.6	44.9	75.9	54.8	46.0	76.8
+PL	49.8	40.9	70.9	50.7	41.7	71.7
+verb expansions	48.7	39.4	69.8	50.1	40.3	70.9
+full LM (40k)	48.2	39.1	68.9	49.2	39.8	69.8
PL (only lexicon)	57.8	47.1	78.8	59.3	47.8	80.7
+verb expansions	56.7	45.9	78.1	57.6	46.5	78.8
+full LM (40k)	53.9	44.0	75.0	56.0	45.7	76.6
English→Spanish						
40k (full corpus)	41.3	34.9	56.1	43.2	35.7	57.8
+PL	40.7	34.4	55.8	42.9	35.9	57.8
1k (small corpus)	57.5	49.2	74.3	58.4	49.9	76.5
+PL	52.4	43.9	68.0	53.6	44.9	68.7
+verb expansions	51.4	43.0	67.7	52.8	44.0	68.4
+full LM (40k)	50.4	42.3	66.2	52.1	43.6	67.6
PL (only lexicon)	61.3	51.8	74.6	62.3	52.7	75.7
+verb expansions	58.1	49.7	73.2	59.5	50.4	74.7
+full LM (40k)	57.6	49.4	72.5	59.1	50.1	73.9

Table 3. Translation Error Rates [%] for the language pair Catalan-English

Catalan→English	Development			Test		
Training Corpus	WER	PER	1-BLEU	WER	PER	1-BLEU
40k (full training)	39.7	32.5	59.8	41.4	33.6	61.1
+PL	38.9	31.8	58.6	41.1	33.2	60.2
1k (reduced training)	53.4	44.8	76.4	54.2	45.0	76.6
+PL	49.2	40.4	71.8	50.9	41.4	72.7
+verb expansions	48.9	39.3	70.9	50.1	39.7	71.4
+full LM (40k)	48.1	38.7	69.8	49.4	39.3	70.4
PL (only lexicon)	57.2	47.3	79.2	59.0	47.9	80.2
+verb expansions	55.0	44.6	76.9	56.1	45.1	76.7
+full LM (40k)	54.0	44.0	75.2	55.6	44.7	75.9
English→Catalan						
40k (full training)	41.5	35.5	58.3	43.3	36.3	60.0
+PL	41.0	35.0	57.8	43.3	36.3	60.1
1k (reduced training)	57.2	49.3	75.2	57.9	49.8	75.1
+PL	52.3	44.2	69.8	53.2	44.9	69.4
+verb expansions	51.6	43.7	69.3	53.0	44.8	70.0
+full LM (40k)	49.7	42.0	66.2	51.2	43.0	67.2
PL (only lexicon)	63.0	53.8	76.3	65.0	55.1	78.2
+verb expansions	58.5	49.8	73.6	59.8	50.6	74.6
+full LM (40k)	57.7	49.3	73.4	59.1	50.2	74.6

Table 4. Translation Error Rates [%] for the external test corpus

Spanish→English	WER	PER	1-BLEU
40k	61.6	49.4	79.1
1k	67.3	56.1	83.1
+PL+exp+lm40k	61.6	49.1	76.7
PL	69.9	57.2	87.6
+exp+lm40k	66.5	54.5	82.0
English→Spanish			
40k	71.3	58.7	76.8
1k	78.8	67.3	84.0
+PL+exp+lm40k	72.5	60.7	77.5
PL	78.8	66.6	84.0
+exp+lm40k	75.2	63.4	79.1
Catalan→English			
40k	64.3	51.5	80.5
1k	71.8	60.5	88.0
+PL+exp+lm40k	68.9	55.7	85.9
PL	73.5	59.7	89.1
+exp+lm40k	70.1	57.0	83.7
English→Catalan			
40k	70.4	58.9	79.8
1k	78.3	67.6	86.8
+PL+exp+lm40k	74.4	62.7	83.1
PL	79.8	68.9	88.0
+exp+lm40k	75.1	64.2	80.6

Table 5. Translation examples English → Spanish using the Phrasal Lexicon (PL) as training data without and with additional verb expansions

source sentence	well I am pretty interested .
PL	bien estoy bastante *interesa* .
PL with expansions	bien estoy bastante *interesado* .
source sentence	there is no problem .
PL	*hay no es* problema
PL with expansions	*no hay* problema .
source sentence	I do not know , what kind of sport do you like best ?
phrasal lexicon	*no me sabe* , qué tipo de deporte te gusta mejor ?
phrasal lexicon with expansions	*no lo sé* , qué tipo de deporte te gusta mejor ?

L. Padró, R. Placer, H. Rodríguez, M. Taulée, and J. Turmo. 1998. An environment for morphosyntactic processing of unrestricted Spanish text. In *Proc. of the first Int. Conf. on Language Resources and Evaluation (LREC)*, Granada, Spain.

E. Matusov, M. Popović, R. Zens, and H. Ney. 2004. Statistical Machine Translation of Spontaneous Speech with Scarce Resources. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 139–146, Kyoto, Japan, September.

S. Nießen and H. Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2):181–204

Table 6. Translation examples Spanish → English using the Phrasal Lexicon (PL) as training data without and with additional verb expansions

source sentence	si , esto , cuántas personas serán ?
PL	if , this , how many people *UNKNOWN_serán* ?
PL with expansions	if , that , how many people *will be* ?
source sentence	queríamos un poco de fruta , yogur , este tipo de cosas , algunas galletas .
PL	*UNKNOWN_queríamos* a little fruit , yogurt , this kind of things , some salted .
PL with expansions	*we wanted* a little fruit , yogurt , this kind of things , some salted .
source sentence	sí , los hoteles , nos movemos entre las tres y las cinco estrellas .
PL	yes , hotels , we *UNKNOWN_movemos* between the three and the five-star .
PL with expansions	yes , the hotels , we *are moving* within the three and the five-star .

F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics*, pages 311–318, Philadelphia, PA, July.

S. Vogel and C. Monson. 2004. Augmenting Manual Dictionaries for Statistical Machine Translation Systems. In *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC)*, pages 1589–1592, Lisbon, Portugal, May.