Linguistic tuple segmentation in ngram-based statistical machine translation

Adrià de Gispert, José B. Mariño

TALP Research Center
Universitat Politècnica de Catalunya (UPC), Barcelona
{agispert|canton}@gps.tsc.upc.edu

Abstract

Ngram-based Statistical Machine Translation relies on a standard Ngram language model of tuples to estimate the translation process. In training, this translation model requires a segmentation of each parallel sentence, which involves taking a hard decision on tuple segmentation when a word is not linked during word alignment. This is especially critical when this word appears in the target language, as this hard decision is compulsory.

In this paper we present a thorough study of this situation, comparing for the first time each of the proposed techniques in two independent tasks, namely English–Spanish European Parliament Proceedings large-vocabulary task and Arabic–English Basic Travel Expressions small-data task. In the face of this comparison, we present a novel segmentation technique which incorporates linguistic information. Results obtained in both tasks outperform all previous techniques.

Index Terms: statistical machine translation, tuple segmentation, n-gram-based SMT, linguistic information

1. Introduction

Ngram-based (or Tuple-based) Statistical Machine Translation (SMT) has proved to be a state-of-the-art alternative approach to phrase-based models in performance comparisons [1, 2]. Its main distinctive treat is the estimation of the core translation model by means of a standard Ngram language model, defined on the special bilingual language expressed by tuples [3].

According to literature, tuples are units containing *one or more* source-language words and *one or more* target-language words (including the NULL token, which is in fact no word). This model has its origins in machine translation using Finite-State Transducers [4], whose theoretical foundation expresses that we seek the target sentence 'e' maximising:

$$\prod_{n=1}^{N} p((f,e)_n|(f,e)_{n-x+1},...,(f,e)_{n-1}) \quad (1)$$

where the n-*th* tuple of a sentence pair is referred as $(f,e)_n$. In order to estimate this language model parameters, given a parallel corpus and a certain word alignment, and in contrast to phrase-based approaches, the training requires a *unique* segmentation of each bilingual pair of sentences into a sequence of tuples so that the natural order of both languages is not violated [5].

The standard tuple extraction algorithm [5] defines a unique set of tuples except whenever the resulting tuple contains no source word (NULL-source tuple). In order to re-use these units in decoding new sentences, the search should allow for no input word to generate units, and this is not the case. Therefore, these units cannot be allowed, and a certain hard decision must be taken regarding tuple segmentation, as in figure 1.

Literature offers examples of criteria to decide this segmentation, ranging from simply linking the target words of the NULL-source tuple to the previous or next tuple (if there is any) deterministically, to comparing the IBM model 1 score assigned to the resulting tuples of two competing segmentations [6].

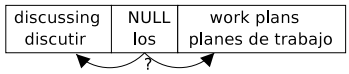

Figure 1: A hard segmentation decision must be taken

However, to our knowledge the impact of this segmentation decision on translation quality is not studied, nor the alternative segmentation strategies are compared. Here, we carry out this comparison, exploring the impact of this hard decision in Ngram translation model estimation and in translation quality, and we propose a novel strategy which follows a linguistic criteria to indirectly reduce the model entropy. We perform this comparison in an English-to-Spanish (and viceversa) large-vocabulary task, and additionally in an Arabic-to-English small-vocabulary task. Furthermore, we study the effect on translation of NULL tokens when these occur in the *target* side of the tuple.

2. Tuple Segmentation strategies

According to the aforementioned conceptual framework of Ngram translation model, it seems clear that the ideal tuple segmentation strategy should take a global decision based on the segmentation for all other NULL-source cases, attempting to obtain that set of tuples and Ngrams which better represented the unseen universe of events, meaning the one with less entropy. However, no feasible algorithm can perform that calculation in a reasonable time given current computational power, as this would involve a whole model reestimation for each particular segmentation alternative.

So far two segmentation strategies to solve the source NULL problem have been proposed, which are presented in the next sections, together with a novel strategy based on a linguistic criterion.

2.1. Deterministic always NEXT

A very pragmatic and simple approach is to always join the target words involved in NULL links (NULL-linked words) to the next tuple, if there is any (otherwise to the previous one), as first done in [7]. Besides simplicity and extreme efficiency, we do not observe any other advantage of this approach, which on the other hand does not follow any linguistic or statistical criterion.

2.2. IBM model 1 weight

Being independent of word position, IBM model 1 probabilities provide a probabilistic lexicon between pairs of word of each language (see [8] for details in these models). This information can be used to weight and compare the resulting tuples from two competing segmentations, is in [6].

While this approach is appealing in that it uses bilingual information, observation of these situations leads to a different conclusion; many NULL-linked words represent articles, prepositions, conjunctions and other particles whose main function is to ensure the grammatical correctness of a sentence, complementing other more informative words. Therefore, their probabilities to translate to another word are not very meaningful.

2.3. Entropy of the POS distribution

Alternatively, from a linguistic point of view, one can think of this tuple segmentation problem around source NULLs as a monolingual decision related to whether the given target word is more connected to the preceding or to the following word. Intuitively, we can expect that a good criterion to perform tuple segmentation lays in preserving grammatically-connected phrases (such as, for instance, articles together with the noun they precede) in the same tuple, as this may probably lead to a simplification of the translation task. On the contrary, splitting linguistic units into separate tuples will probably lead to a tuple vocabulary increase and a higher sparseness, producing a poorer Ngram model.

In this direction, we propose to take the segmentation decision according to the entropy of the forward and backward Part-Of-Speech (POS) distributions, which we define conditioned to context. In detail, given the tuple sequence described as follows:

$$\begin{array}{ccc} < ...s_j > & \text{NULL} & < s_{j+1}... > \\ | & | & | \\ < ...t_{i-1} > & t_i & < t_{i+1}... > \end{array}$$

where s_j means word in position j in source sentence, and equivalently t_i means word in position i in target sentence, we can define a 'forward' entropy of the POS distribution in position $i+1$ given (t_{i-1}, t_i) as in equation 2:

$$H^f_{POS} = -\sum_{POS} p^f_{POS} \log p^f_{POS} \quad (2)$$

where

$$p^f_{POS} = \frac{N(t_{i-1}, t_i, POS_{i+1})}{\sum_{POS'} N(t_{i-1}, t_i, POS'_{i+1})} \quad (3)$$

is the probability of observing a certain Part-Of-Speech *following* the sequence of words defined by t_i and t_{i+1}.

Equivalently, we can define a 'backward' entropy of the POS distribution in position $i-1$ given (t_i, t_{i+1}) as in equation 4:

$$H^b_{POS} = -\sum_{POS} p^b_{POS} \log p^b_{POS} \quad (4)$$

where

$$p^b_{POS} = \frac{N(POS_{i-1}, t_i, t_{i+1})}{\sum_{POS'} N(POS'_{i-1}, t_i, t_{i+1})} \quad (5)$$

is the probability of observing a certain Part-Of-Speech *preceding* the sequence of words defined by t_{i-1} and t_i.

Then, we can take a tuple segmentation decision favouring the most POS-entropic case. The rationale behind this is that, if $H^f_{POS} > H^b_{POS}$, we have observed the first sequence of words comprised of (t_{i-1}, t_i) in more grammatically different situations than the latter sequence comprised of (t_i, t_{i+1}). Therefore, we can induce that t_{i-1} and t_i tend to be more often connected than t_i and t_{i+1}, and should belong to the same translation tuple. Analogously, one can conclude the contrary if $H^f_{POS} < H^b_{POS}$.

While this is a monolingual decision on the target language morphology, being linguistically-guided, the POS entropy approach is much more correlated with human intuition.

2.4. Criteria for NULLs in target

Whereas the segmentation decision is required when a target word is unlinked (or linked to NULL), this is not so when the unlinked word is in the source target, in which case these units are allowed in the tuple vocabulary for Ngram estimation .

However, one can think of applying the same criterion to remove NULLs in the target side of tuples, possibly addressing omission errors in translation. Aiming at evaluating the impact of this decision, we have also applied the POS entropy strategy to segment tuples with unlinked source words.

3. Experimental framework

In order to compare each segmentation strategy and evaluate its impact on translation quality, experiments have been carried out using two parallel corpora, differing in language pair and corpus size. On the one hand, we used a Spanish–English large-vocabulary corpus, containing the European Parliament Proceedings from 1996 to September 2004, and on the other, an Arabic-to-English small-vocabulary corpus, which contains a small part of the Basic Travel Expressions Corpus. English has been tagged using *TnT* tagger[1], and Spanish using *FreeLing* analysis tool[2].

	sent	words	vocab	avglen	refs
Train set					
English	1.22 M	33.37 M	104.8 k	27.3	1
Spanish		34.96 M	151.4 k	28.6	
Dev set				OOVs	
English	504	15.3 k	2.30 k	20	3
Spanish		15.4 k	2.74 k	22	3
Test set				OOVs	
English	840	22.75 k	4.1 k	44	2
Spanish	1094	26.88 k	4.0 k	113	2

Table 1: *European Parliament English-Spanish corpus statistics.*

Statistics of each corpus can be found in Tables 1 and 2, including number of sentences, running words, vocabulary sizes, average sentence length and out-of-vocabulary (OOVs) words and number of reference translations for development and test sets. Note the big size difference between tasks.

3.1. Translation model results

A comparison of Ngram translation model performance for each task is shown in rows named 'alwaysNEXT', 'IBM1weight' and

[1] Available at www.coli.uni-saarland.de/∼thorsten/tnt

[2] Available at http://garraf.epsevg.upc.es/freeling

	sent	words	vocab	avglen	refs
Train set					
Arabic	20 k	180.5 k	16.0 k	9.0	1
English		189.2 k	7.2 k	9.5	
Dev set				OOVs	
Arabic	506	3.63 k	1.18 k	196	16
Test set				OOVs	
Arabic	1006	7.22 k	1.9 k	356	16

Table 2: *Travel Expressions Arabic-English corpus statistics.*

		BLEU	mWER	NIST
E→S	alwaysNEXT	0.4215	43.98	9.22
	IBM1weight	0.4221	43.60	9.19
	POSentropy	**0.4325**	**43.48**	**9.30**
	trgNULL	0.4249	44.47	9.21
	trgNULLpos	0.4313	43.75	9.29
S→E	alwaysNEXT	0.4661	39.37	9.86
	IBM1weight	0.4698	38.73	9.91
	POSentropy	**0.4756**	**38.64**	**9.95**
	trgNULL	0.4728	39.23	9.91
	trgNULLpos	0.4733	38.78	9.93
A→E	alwaysNEXT	0.3684	41.80	7.16
	IBM1weight	0.3656	41.94	7.14
	POSentropy	**0.3691**	41.91	**7.17**

Table 3: *Translation model performance for each segmentation strategy. 'E' stands for English, 'S' for Spanish and 'A' for Arabic.*

'POSentropy' in Table 3, refering to each segmentation strategy discussed in sections 2.1 to 2.3, respectively.

Regarding the large-vocabulary tasks, the proposed linguistically-guided segmentation outperforms all other strategies significantly, especially in the Eng→Spa direction. This result is consistent with the fact that Spanish is a more word-generative language than English, and therefore, more NULLs are found in the English side of extracted tuples.

Even though the impact of changing the segmentation criterion when translating into English is smaller, the improvement of the POSentropy approach is significant. In the small-vocabulary Ara→Eng task differences are less significant, in correlation with the fact that *only* 7% of tuples contain NULLs in Arabic side, compared to the 14% of Eng→Spa task.

Remarkably, whereas IBM1weight provides better results in large-vocabulary tasks than the alwaysNEXT criterion, the result is opposite in the small-vocabulary Ara→Eng task. On the other hand, the POSentropy approach proves to be more general and robust to a task change, achieving best performance in all tasks.

3.2. Removing NULLs in target

Following the idea suggested in section 2.4, Table 3 also presents results when applying the best segmentation criterion (POSentropy) to avoid NULLs in the tuples *target* side, as shown in rows named 'trgNULL' and 'trgNULLpos' for large-vocabulary Spanish–English task. The first refers to applying the criterion to all tuples, whereas the latter to only applying it when the tuple contains a POS of a Noun, Adjective or Verb. The objective of this is to minimise omission errors by preventing tuples with content words in source side and NULL in target to belong to the model dictionary.

However, results show that none of these techniques is beneficial. Clearly, and in contrast to NULLs in the tuple source side, NULLs in the target side are a useful mechanism for the Ngram model to find good contexts and significantly increase performance regardless of the translation direction. This holds even when we allow tuples with content words in source side and NULL in target.

3.3. Translation Ngrams study

To better understand these results, Table 4 shows the tuple vocabulary obtained in training for each segmentation (tup vcb), and relevant statistics of translated output, namely the percentage of tuples used as 1-grams, 2-grams and 3-grams during translation, the average tuple length obtained (measuring source and target sides separately) and the number of tuples with NULL in target (in the translated output).

Regarding tuple vocabulary size, the alwaysNEXT criterion produces the biggest vocabulary in training when compared to POSentropy and IBM1weight, which produces the smallest. When removing the target NULLs, the vocabulary size is significantly improved.

In Eng→Spa, we observe that translation with alwaysNEXT and POSentropy segmentation criteria tend to use more 3grams than IBM1weight, which can be explained by their consistency in taking segmentation decisions (they invariably take the same decision given the target words involved), whereas IBM1weight depends on source and target words and is more variable.

However, using more 3grams is not directly correlated with translation scores, and the number of tuples to NULL needs to be taken into account. The high number of tuples to NULL for the alwaysNEXT criterion is outstanding, and tells us that translation is indeed achieving many 3grams by catenating sequences to NULL, which do not necessarily achieve better performance. In the case of IBM1weight and especially of POSentropy, the number of tuples with NULL in target is strongly reduced. Whereas this appears to be positive for translation performance, when completely or partially removing NULLs in target (trgNULL and trgNULLpos), average tuple length increases, not only in the source side but also in the target side, and the model looses tuple context and falls much more often to 1gram. Apparently, this has a negative effect in translation quality.

Therefore, we can conclude that the best relationship between high-order tuple context and small amount of tuples to NULL is achieved by the proposed POSentropy segmentation criterion. Differences are much smaller in the Spa→Eng direction, although the same tendency in number of tuples with target NULL is to be found, and conclusions are analogous.

3.4. Translation model + features results

To further evaluate the impact of these segmentation differences, we have log-linearly combined the Ngram translation model with additional features, namely two lexicon features based on IBM model 1 probabilities, a standard target 3-gram Language Model and a constant word bonus, and optimized according to BLEU score in development set, as similarly done in [3]. Table 5 shows the translation results for the two best segmentations for each task. As it can be seen, the improvement of better segmenting tuples with NULLs in source is practically compensated by the contribution of additional features, especially in the Spa→Eng task.

	Eng→Spa				Spa→Eng			
	tup vcb	% 1–2–3grams	tup len	null	tup vcb	% 1–2–3grams	tup len	null
alwaysNEXT	2.11 M	17.6 – 44.4 – 38.0	1.16-1.10	3119	2.15 M	14.1 – 41.5 – 44.4	1.14-1.06	2761
IBM1weight	2.04 M	18.0 – 44.7 – 37.3	1.16-1.09	2466	2.08 M	14.2 – 41.4 – 44.4	1.13-1.05	2318
POSentropy	2.08 M	17.8 – 44.3 – 37.9	1.16-1.11	2282	2.11 M	14.2 – 41.5 – 44.3	1.13-1.06	2194
trgNULL	2.35 M	23.2 – 45.1 – 31.7	1.25-1.19	0	2.42 M	19.9 – 44.1 – 36.0	1.26-1.22	0
trgNULLpos	2.18 M	19.0 – 44.5 – 36.5	1.18-1.14	1625	2.16 M	14.7 – 41.6 – 43.7	1.14-1.08	1977

Table 4: *Tuple vocabulary and Ngram translation statistics for each segmentation strategy.*

		BLEU	mWER	NIST
E→S	IBM1weight	0.4714	40.22	9.83
	POSentropy	**0.4744**	40.56	**9.85**
S→E	IBM1weight	0.5470	34.41	10.74
	POSentropy	0.5466	34.44	10.72
A→E	alwaysNEXT	0.3974	40.16	7.23
	POSentropy	**0.4024**	**40.05**	**7.39**

Table 5: *Results with additional features.*

In the large-vocabulary Spa-Eng tasks the target language and lexicon models add robustness to the whole system by penalising tuples with 'wrong' segmentations, or at least their catenation to build up the translation output. Yet the proposed segmentation achieves slightly better results in the Eng→Spa direction.

However, small-vocabulary tasks seem more sensitive to segmentation even when combining the core translation model with additional features, as improvements are more significant and higher than using only the translation model.

4. Conclusion

This paper delves into the details of tuple segmentation, a necessary step when training Ngram-based SMT systems. Apart from comparing all previously presented segmentation criteria, a novel strategy based on the entropy of the POS distribution is proposed.

A first conclusion of our study is that translation model performance is significantly affected by tuple segmentation, and the impact of segmentation strategies depends on language pair and corpus used, growing according to the percentage of tuples with NULL in source side. Secondly, the proposed linguistic criterion performs significantly better than segmentation strategies already presented in literature, behaving consistently across different translation tasks.

Regarding NULLs in target side, we conclude that they provide context to the model and removing them through resegmentation is not beneficial. Finally, when translation model is loglinearly combined with other features, direct impact of segmentation is smaller for large-vocabulary tasks, as features can partially compensate a bad tuple segmentation.

Possibly, the main drawback of the proposed segmentation strategy is that it requires a POS tagging tool. In this direction, one can think of making it independent from this by using induced classification strategies. Opposite to that, if chunking tools are available (for example in English), we would like to investigate ways of using chunking information to better segment tuples.

Another interesting issue is related to NULLs in target side. Whereas the presented experiments partially or fully removing NULLs in target did not yield improvements, it appears that the best performing strategy is the one that 'uses' less units to NULL in translation. A further study of the reasons behind this is to be carried out, trying to understand the real contribution of this units and how or whether further improvements could be achieved by removing some of them.

Finally, other issues related to the tuple ngram translation model are pruning strategies and smoothing techniques. Though out of the scope of this paper, they are also crucial to understand the core model of ngram-based SMT. Therefore, a thorough study of their respective impact in translation quality is to be carried out.

5. Acknowledgements

This work was partly funded by European Union (TC-STAR project, FP6-506738), Generalitat de Catalunya and European Social Fund.

6. References

[1] P. Koehn and C. Monz, "Shared task: Statistical Machine Translation between European Languages," *Proc. of the ACL Workshop on Building and Using Parallel Texts (ACL'05)*, pp. 119–124, 2005.

[2] M. Eck and C. Hori, "Overview of the IWSLT 2005 Evaluation Campaign," *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, pp. 11–32, 2005.

[3] J. Mariño, R. Banchs, J. Crego, A. de Gispert, P. Lambert, M. R. Costa-jussà, and J. Fonollosa, "Bilingual N–gram statistical machine translation," *Proc. of the MT Summit X*, pp. 275–282, 2005.

[4] E. Vidal, "Finite-state speech-to-speech translation," *Proc. of 1997 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 111–114, 1997.

[5] J. M. Crego, J. Mariño, and A. de Gispert, "Finite-state-based and phrase-based statistical machine translation," *Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04*, pp. 37–40, 2004.

[6] J. M. Crego, J. Mariño, and A. Gispert, "TALP: The UPC tuple-based SMT system," *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, pp. 191–198, 2005.

[7] A. de Gispert and J. Mariño, "TALP: Xgram-based Spoken Language Translation System," *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pp. 85–90, 2004.

[8] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.