Can We Translate Letters?

David Vilar, Jan-T. Peter and Hermann Ney
Lehrstuhl für Informatik 6
RWTH Aachen University
D-52056 Aachen, Germany
{vilar,peter,ney}@cs.rwth-aachen.de

Abstract

Current statistical machine translation systems handle the translation process as the transformation of a string of symbols into another string of symbols. Normally the symbols dealt with are the words in different languages, sometimes with some additional information included, like morphological data. In this work we try to push the approach to the limit, working not on the level of words, but treating both the source and target sentences as a string of letters. We try to find out if a nearly unmodified state-of-the-art translation system is able to cope with the problem and whether it is capable to further generalize translation rules, for example at the level of word suffixes and translation of unseen words. Experiments are carried out for the translation of Catalan to Spanish.

1 Introduction

Most current statistical machine translation systems handle the translation process as a "blind" transformation of a sequence of symbols, which represent the words in a source language, to another sequence of symbols, which represent words in a target language. This approach allows for a relative simplicity of the models, but also has drawbacks, as related word forms, like different verb tenses or plural-singular word pairs, are treated as completely different entities.

Some efforts have been made e.g. to integrate more information about the words in the form of Part Of Speech tags (Popović and Ney, 2005), using additional information about stems and suffixes (Popović and Ney, 2004) or to reduce the morphological variability of the words (de Gispert, 2006). State of the art decoders provide the ability of handling different word forms directly in what has been called factored translation models (Shen et al., 2006).

In this work, we try to go a step further and treat the words (and thus whole sentences) as sequences of letters, which have to be translated into a new sequence of letters. We try to find out if the translation models can generalize and generate correct words out of the stream of letters. For this approach to work we need to translate between two related languages, in which a correspondence between the structure of the words can be found.

For this experiment we chose a Catalan-Spanish corpus. Catalan is a romance language spoken in the north-east of Spain and Andorra and is considered by some authors as a transitional language between the Iberian Romance languages (e.g. Spanish) and Gallo-Romance languages (e.g. French). A common origin and geographic proximity result in a similarity between Spanish and Catalan, albeit with enough differences to be considered different languages. In particular, the sentence structure is quite similar in both languages and many times a nearly monotonical word to word correspondence between sentences can be found. An example of Catalan and Spanish sentences is given in Figure 1.

The structure of the paper is as follows: In Section 2 we review the statistical approach to machine translation and consider how the usual techniques can be adapted to the letter translation task. In Sec-

Proceedings of the Second Workshop on Statistical Machine Translation, pages 33–39,
Prague, June 2007.

Catalan	Perquè a mi m'agradaria estar-hi dues, una o dues setmanes, més o menys, depenent del preu i cada hotel.
Spanish	Porque a mí me gustaría quedarme dos, una o dos semanas, más o menos, dependiendo del precio y cada hotel.
English	Because I would like to be there two, one or two weeks, more or less, depending on the price of each hotel.

Catalan	Si baixa aquí tenim una guia de la ciutat que li podem facilitar en la que surt informació sobre els llocs més interessants de la ciutat.
Spanish	Si baja aquí tenemos una guía de la ciudad que le podemos facilitar en la que sale información sobre los sitios más interesantes de la ciudad.
English	If you come down here we have a guide book of the city that you can use, in there is information about the most interesting places in the city.

Figure 1: Example Spanish and Catalan sentences (the English translation is provided for clarity).

tion 3 we present the results of the letter-based translation and show how to use it for improving translation quality. Although the interest of this work is more academical, in Section 4 we discuss possible practical applications for this approach. The paper concludes in Section 5.

2 From Words To Letters

In the standard approach to statistical machine translation we are given a sentence (sequence of words) $f_1^J = f_1 \dots f_J$ in a source language which is to be translated into a sentence $\hat{e}_1^I = \hat{e}_1 \dots \hat{e}_I$ in a target language. Bayes decision rule states that we should choose the sentence which maximizes the posterior probability

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} p(e_1^I|f_1^J)\,, \tag{1}$$

where the argmax operator denotes the search process. In the original work (Brown et al., 1993) the posterior probability $p(e_1^I|f_1^J)$ is decomposed following a noisy-channel approach, but current state-of-the-art systems model the translation probability directly using a log-linear model(Och and Ney, 2002):

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)}\,, \tag{2}$$

with h_m different models, λ_m scaling factors and the denominator a normalization factor that can be ignored in the maximization process. The λ_m are usually chosen by optimizing a performance measure over a development corpus using a numerical optimization algorithm like the downhill simplex algorithm (Press et al., 2002).

The most widely used models in the log linear combination are phrase-based models in source-to-target and target-to-source directions, ibm1-like scores computed at phrase level, also in source-to-target and target-to-source directions, a target language model and different penalties, like phrase penalty and word penalty.

This same approach can be directly adapted to the letter-based translation framework. In this case we are given a sequence of letters $\mathcal{F}_1^{\mathcal{J}}$ corresponding to a source (word) string f_1^J, which is to be translated into a sequence of letters $\mathcal{E}_1^{\mathcal{I}}$ corresponding to a string e_1^I in a target language. Note that in this case whitespaces are also part of the vocabulary and have to be generated as any other letter. It is also important to remark that, without any further restrictions, the word sequences e_1^I corresponding to a generated letter sequence $\mathcal{E}_1^I$ are not even composed of actual words.

2.1 Details of the Letter-Based System

The vocabulary of the letter-based translation system is some orders of magnitude smaller than the vocabulary of a full word-based translation system, at least for European languages. A typical vocabulary size for a letter-based system would be around 70, considering upper- and lowercase letter, digits, whitespace and punctuation marks, while the vocabulary size of a word-based system like the ones used in current evaluation campaigns is in the range of tens or hundreds of thousands words. In a normal situation there are no unknowns when carrying out the actual translation of a given test corpus. The situation can be very different if we consider languages like Chinese or Japanese.

This small vocabulary size allows us to deal with a larger context in the models used. For the phrase-based models we extract all phrases that can be used when translating a given test corpus, without any restriction on the length of the source or the target part[1]. For the language model we were able to use a high-order n-gram model. In fact in our experiments a 16-gram letter-based language model is used, while state-of-the-art translation systems normally use 3 or 4-grams (word-based).

In order to better try to generate "actual words" in the letter-based system, a new model was added in the log-linear combination, namely the count of words generated that have been seen in the training corpus, normalized with the length of the input sentence. Note however that this models enters as an additional feature function in the model and it does not constitute a restriction of the generalization capabilities the model can have in creating "new words". Somehow surprisingly, an additional word language model did not help.

While the vocabulary size is reduced, the average sentence length increases, as we consider each letter to be a unit by itself. This has a negative impact in the running time of the actual implementation of the algorithms, specially for the alignment process. In order to alleviate this, the alignment process was split into two passes. In the first part, a word alignment was computed (using the GIZA++ toolkit (Och and Ney, 2003)). Then the training sentences were split according to this alignment (in a similar way to the standard phrase extraction algorithm), so that the length of the source and target part is around thirty letters. Then, a letter-based alignment is computed.

2.2 Efficiency Issues

Somewhat counter-intuitively, the reduced vocabulary size does not necessarily imply a reduced memory footprint, at least not without a dedicated program optimization. As in a sensible implementations of nearly all natural language processing tools, the words are mapped to integers and handled as such. A typical implementation of a phrase table is then a prefix-tree, which is accessed through these word indices. In the case of the letter-based translation, the phrases extracted are much larger than the word-based ones, in terms of elements. Thus the total size of the phrase table increases.

The size of the search graph is also larger for the letter-based system. In most current systems the generation algorithm is a beam search algorithm with a "source synchronous" search organization. As the length of the source sentence is dramatically increased when considering letters instead of words, the total size of the search graph is also increased, as is the running time of the translation process.

The memory usage for the letter system can actually be optimized, in the sense that the letters can act as "indices" themselves for addressing the phrase table and the auxiliary mapping structure is not necessary any more. Furthermore the characters can be stored in only one byte, which provides a significant memory gain over the word based system where normally four bytes are used for storing the indices. These gains however are not expected to counteract the other issues presented in this section.

3 Experimental Results

The corpus used for our experiment was built in the framework of the LC-STAR project (Conejero et al., 2003). It consists of spontaneous dialogues in Spanish, Catalan and English[2] in the tourism and travelling domain. The test corpus (and an additional development corpus for parameter optimization) was randomly extracted, the rest of the sentences were used as training data. Statistics for the corpus can be seen in Table 1. Details of the translation system used can be found in (Mauser et al., 2006).

The results of the word-based and letter-based approaches can be seen in Table 2 (rows with label "Full Corpus"). The high BLEU scores (up to nearly 80%) denote that the quality of the translation is quite good for both systems. The word-

[1] For the word-based system this is also the case.

[2] The English part of the corpus was not used in our experiments.

		Spanish	Catalan
Training	Sentences	40 574	
	Running Words	482 290	485 514
	Vocabulary	14 327	12 772
	Singletons	6 743	5 930
Test	Sentences	972	
	Running Words	12 771	12 973
	OOVs [%]	1.4	1.3

Table 1: Corpus Statistics

based system outperforms the letter-based one, as expected, but the letter-based system also achieves quite a good translation quality. Example translations for both systems can be found in Figure 2. It can be observed that most of the words generated by the letter based system are correct words, and in many cases the "false" words that the system generates are very close to actual words (e.g. "elos" instead of "los" in the second example of Figure 2).

We also investigated the generalization capabilities of both systems under scarce training data conditions. It was expected that the greater flexibility of the letter-based system would provide an advantage of the approach when compared to the word-based approach. We randomly selected subsets of the training corpus of different sizes ranging from 1 000 sentences to 40 000 (i.e. the full corpus) and computed the translation quality on the same test corpus as before. Contrary to our hopes, however, the difference in BLEU score between the word-based and the letter-based system remained fairly constant, as can be seen in Figure 3, and Table 2 for representative training corpus sizes.

Nevertheless, the second example in Figure 2 provides an interesting insight into one of the possible practical applications of this approach. In the example translation of the word-based system, the word "centreamericans" was not known to the system (and has been explicitly marked as unknown in Figure 2). The letter-based system, however, was able to correctly learn the translation from "centre-" to "centro-" and that the ending "-ans" in Catalan is often translated as "-anos" in Spanish, and thus a correct translation has been found. We thus chose to combine both systems, the word-based system doing most of the translation work, but using the letter-based system for the translation of unknown words. The results of this combined approach can be found in Table 2 under the label "Combined System". The combination of both approaches leads to a 0.5% increase in BLEU using the full corpus as training material. This increase is not very big, but is it over a quite strong baseline and the percentage of out-of-vocabulary words in this corpus is around 1% of the total words (see Table 1). When the corpus size is reduced, the gain in BLEU score becomes more important, and for the small corpus size of 1 000 sentences the gain is 2.5% BLEU. Table 2 and Figure 3 show more details.

4 Practical Applications

The approach described in this paper is mainly of academical interest. We have shown that letter-based translation is in principle possible between similar languages, in our case between Catalan and Spanish, but can be applied to other closely related language pairs like Spanish and Portuguese or German and Dutch. The approach can be interesting for languages where very few parallel training data is available.

The idea of translating unknown words in a letter-based fashion can also have applications to state-of-the-art translation systems. Nowadays most automatic translation projects and evaluations deal with translation from Chinese or Arabic to English. For these language pairs the translation of named entities poses an additional problem, as many times they were not previously seen in the training data and they are actually one of the most informative words in the texts. The "translation" of these entities is in most cases actually a (more or less phonetic) transliteration, see for example (Al-Onaizan and Knight, 2002). Using the proposed approach for the translation of these words can provide a tighter integration in the translation process and hopefully increase the translation performance, in the same way as it helps for the case of the Catalan-Spanish translation for unseen words.

Somewhat related to this problem, we can find an additional application in the field of speech recognition. The task of grapheme-to-phoneme conversion aims at increasing the vocabulary an ASR system can recognize, without the need for additional

		BLEU	WER	PER
Word-Based System	Full Corpus	78.9	11.4	10.6
	10k	74.0	13.9	13.2
	1k	60.0	21.3	20.1
Letter-Based System	Full Corpus	72.9	14.7	13.5
	10k	69.8	16.5	15.1
	1k	55.8	24.3	22.8
Combined System	Full Corpus	79.4	11.2	10.4
	10k	75.2	13.4	12.6
	1k	62.5	20.2	19.0

Table 2: Translation results for selected corpus sizes. All measures are percentages.

Source (Cat)	Bé, en principi seria per a les vacances de Setmana Santa que són les següents que tenim ara, entrant a juliol.
Word-Based	Bueno, en principio sería para las vacaciones de Semana Santa que son las siguientes que tenemos ahora, entrando en julio.
Letter-Based	Bueno, en principio sería para las vacaciones de Semana Santa que son las siguientes que tenemos ahora, entrando bamos en julio .
Reference	Bueno, en principio sería para las vacaciones de Semana Santa que son las siguientes que tenemos ahora, entrando julio.

Source (Cat)	Jo li recomanaria per exemple que intentés apropar-se a algun país veí també com poden ser els països centreamericans, una mica més al nord Panamá.
Word-Based	Yo le recomendaría por ejemplo que intentase acercarse a algún país vecino también como pueden ser los países UNKNOWN_centreamericans, un poco más al norte Panamá.
Letter-Based	Yo le recomendaría por ejemplo que intentaseo acercarse a algún país veí también como pueden ser elos países centroamericanos, un poco más al norte Panamá.
Combined	Yo le recomendaría por ejemplo que intentase acercarse a algún país vecino también como pueden ser los países centroamericanos, un poco más al norte Panamá.
Reference	Yo le recomendaría por ejemplo que intentase acercarse a algún país vecino también como pueden ser los países centroamericanos, un poco más al norte Panamá.

Figure 2: Example translations of the different approaches. For the word-based system an unknown word has been explicitly marked.

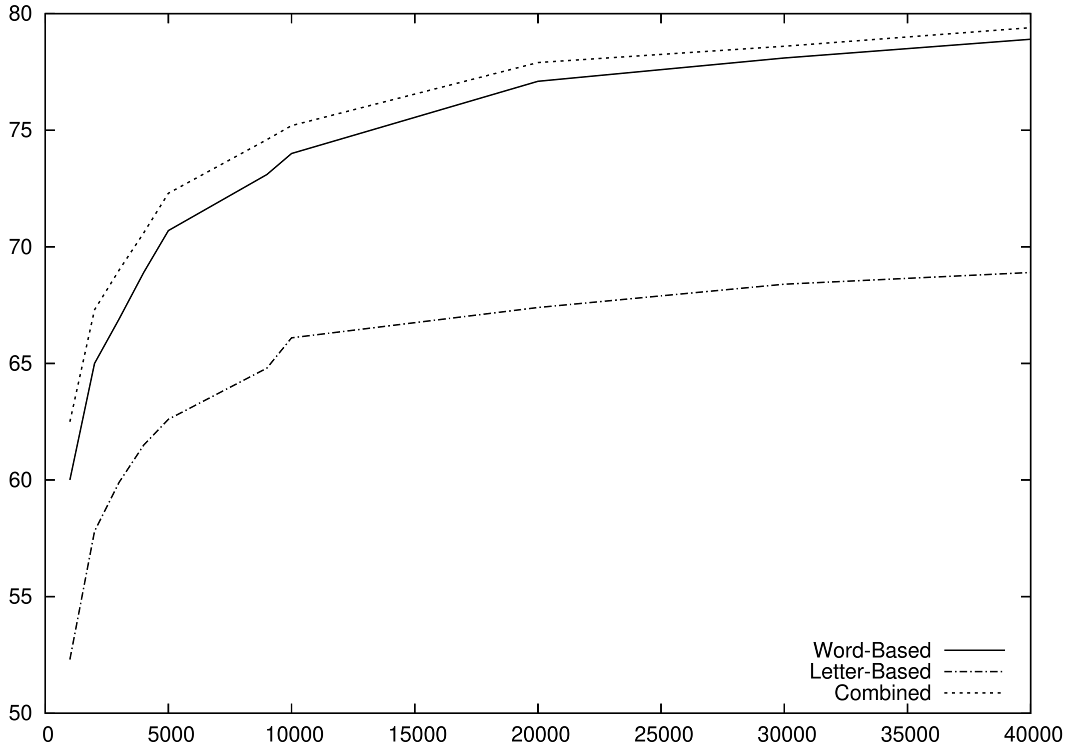

Figure 3: Translation quality depending of the corpus size.

acoustic data. The problem can be formulated as a translation from graphemes ("letters") to a sequence of graphones ("pronunciations"), see for example (Bisani and Ney, 2002). The proposed letter-based approach can also be adapted to this task.

Lastly, a combination of both, word-based and letter-based models, working in parallel and perhaps taking into account additional information like base forms, can be helpful when translating from or into rich inflexional languages, like for example Spanish.

5 Conclusions

We have investigated the possibility of building a letter-based system for translation between related languages. The performance of the approach is quite acceptable, although, as expected, the quality of the word-based approach is superior. The combination of both techniques, however, allows the system to translate words not seen in the training corpus and thus increase the translation quality. The gain is specially important when the training material is scarce.

While the experiments carried out in this work are more interesting from an academical point of view, several practical applications has been discussed and will be the object of future work.

Acknowledgements

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistische Textübersetzung" (NE 572/5-3).

References

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in arabic text. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–13, Morristown, NJ, USA. Association for Computational Linguistics.

Max Bisani and Hermann Ney. 2002. Investigations on joint-multigram models for grapheme-to-phoneme conversion. In *Proceedings of the 7th International Conference on Spoken Language Processing*, volume 1, pages 105–108, Denver, CO, September.

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*, 19(2):263–311, June.

D. Conejero, J. Gimnez, V. Arranz, A. Bonafonte, N. Pascual, N. Castell, and A. Moreno. 2003. Lexica and corpora for speech-to-speech translation: A trilingual approach. In *European Conf. on Speech Communication and Technology*, pages 1593–1596, Geneva, Switzerland, September.

Adrià de Gispert. 2006. *Introducing Linguistic Knowledge into Statistical Machine Translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, October.

Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. 2006. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *Proc. of the International Workshop on Spoken Language Translation*, pages 103–110, Kyoto, Japan.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Maja Popović and Hermann Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *4th International Conference on Language Resources and Evaluation (LREC)*, pages 1585–1588, Lisbon, Portugal, May.

Maja Popović and Hermann Ney. 2005. Exploiting Phrasal Lexica and Additional Morpho-syntactic Language Resources for Statistical Machine Translation with Scarce Training Data. In *10th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 212–218, Budapest, Hungary, May.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

Wade Shen, Richard Zens, Nicola Bertoldi, and Marcello Federico. 2006. The JHU Workshop 2006 IWSLT System. In *Proc. of the International Workshop on Spoken Language Translation*, pages 59–63, Kyoto, Japan.