New reordering and modeling strategies for Statistical Machine Translation

Marta R. Costa-jussà[1]*; Supervisor: José A. R. Fonollosa*[2]

[1] Barcelona Media Innovation Center
marta.ruiz@barcelonamedia.org
[2] Universitat Politècnica de Catalunya
adrian@gps.tsc.upc.edu

Abstract

Nowadays, translation may be the bottleneck of the pretended information globalisation. While surfing the Internet, for instance, sometimes we come across languages and characters we do not understand. Statistical machine translation (SMT) constitutes a research sub-area of machine translation (MT) that has recently gained much popularity. In fact, this technology has experienced real growth motivated by the development of computer resources needed to implement translation algorithms based on statistical methods. This thesis focuses on the SMT framework and primarily on the definition and experimentation of novel algorithms for building a correct structural reordering for translated words. Moreover, challenging techniques regarding language modeling and system combination are successfully applied to state-of-the-art SMT systems. This thesis should shed some light on the SMT approach and on the word ordering challenges and should be specially useful to natural language processing researchers having non or some expertise in machine translation.

Index Terms: Statistical machine translation, Word reordering, Language modeling, System combination, Rescoring, Word graphs

1. Introduction

This thesis focuses on the statistical machine translation (SMT) framework and primarily on the definition and experimentation of novel algorithms for building a correct structural reordering for translated words. Moreover, challenging techniques regarding language modeling and system combination are successfully applied to state-of-the-art SMT systems.

To begin, a thorough study of the SMT state-of-the-art is performed. Ngram- and phrase-based SMT feature functions are described. The former, which has been developed in our research group, is used as a baseline system and the latter, given its popularity, is used to deepen the new techniques during experimentation.

This thesis proposes the introduction of novel statistical reordering techniques in an SMT system. The first approach is based on an algorithm that detects, learns and infers pairs of words in the source language that swap in the target language providing accurate local reorderings. The second approach consists of generating weighted reordering hypotheses using the same powerful techniques of SMT systems in order to undo the source language structure and to make it more similar to the target language structure. Therefore, the translation challenge is divided into two steps: predicting the order of the words in the target language and substituting these words in the target language. In order to infer new reorderings that were not learnt during training, the NbR system uses word classes instead of words themselves. In order to correctly integrate the NbR and SMT systems, both are concatenated, by using a word graph. This approach is an elegant and efficient reordering approach that is capable of achieving significantly improved translation in the target language.

Then, the introduction of continuous space language models is reported and analyzed in an Ngram-based system that uses translation and target language models. The continuous space language modeling technique is based on projecting word indices onto a continuous space. The resulting probability functions are smooth functions of the word representation. Events are better estimated than in standard smoothing methods, which is shown by the significant reduction in perplexity. This better probability estimation allows for an improvement in translation quality.

Moreover, this thesis performs a two-system combination considering the phrase and Ngram-based systems. Multiple outputs of both systems with their corresponding score are concatenated, and for each system translation the score given by the opposite system is computed. The final translation is properly chosen by simultaneously considering the scores given by both systems.

This paper reviews the main thesis ideas and it is organized as follows. Next section describes the nature of the machine translation problem. Section 3 presents the relevant theory from a qualitative point of view. Section 4 explains the thesis goals. Section 5 describes the novel methods proposed in the thesis. Finally, section 6 reports the project framework of this thesis and section 7, the achievements.

2. Nature of the problem

This PhD thesis focuses on the framework of statistical machine translation (SMT), which is a specific approach to machine translation (MT). The main goal of MT is to be able to translate from a source language s to a target language t. MT is a difficult task, mainly because natural languages are highly complex. Many words have more than one meaning and sentences may have various readings. Certain grammatical relations in one language might not exist in another language. Moreover, there are non-linguistic factors such as the problem that performing a translation might require world knowledge. Additional challenges arise when dealing with spoken language translation like confronting non-grammatical texts.

In order to face the MT challenge, many dependencies have to be taken into account. Often, these dependencies are weak and vague, which makes it rarely possible to describe simple and relevant rules that hold without exception for different language pairs. SMT treats MT as a decision problem, where we have to decide upon several target sentences, given a source sentence and among all possible target sentences, we will choose the sentence with the highest probability according to a statistically-learned model. SMT technology has received increasing interest leading to improved algorithms and it has been justified by various successful comparative evaluations since its revival by the work of the famous IBM research group more than fifteen years ago. It has proved to be a competitive approach, which shows greater robustness than other methods for the translation of spontaneous speech. Particularly, SMT translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. However, translations generated by SMT systems still have several significant challenges to pursue, like word reordering or word correspondences as we will see in the next sections.

3. Relevant theory

The SMT framework formulates the problem of translating a sentence from a source language s into a target language t as the maximization problem of the conditional probability $p(t|s)$. During the translation process, a statistical score based on the probabilities of the feature functions is assigned to each translation candidate, and the one with the highest combination score is selected as translation output. However, the SMT system might not be able to correctly score translations due to statistical models limitations.

Reordering is understood as the word order redistribution of the translated words as shown in Figure 1. In initial SMT systems, this different order is only modeled within the limits of translation units.

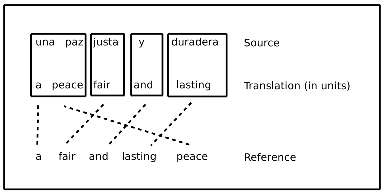

Figure 1: *Source, translation (in units) and reference example. Translation and reference differ in word order.*

Relying only in the reordering provided by translation units may not be good enough in most language pairs, which might require longer reorderings (as shown in Figure 1). Therefore, additional techniques may be deployed to face the reordering challenge. That is why many extended approachespropose to face statistical machine translation as a concatenation of two sub-tasks: predicting the collection of words in a translation and deciding the order of the predicted words. Introducing more complex feature functions that facilitate scoring the translation may not be easily introduced during decoding; for example, those feature functions that use the entire sentence to produce a score. One straightforward solution to this problem is a two-step decoding approach: in the first step, the decoder is run in n-best mode to produce n-best lists with N hypotheses per sentence. In the second step, the n-best lists are rescored with additional models. This technique reevaluates the n-best translation hypotheses of an MT system by introducing additional feature functions that should add information not included during decoding. Figure 2 shows a standard rescoring framework.

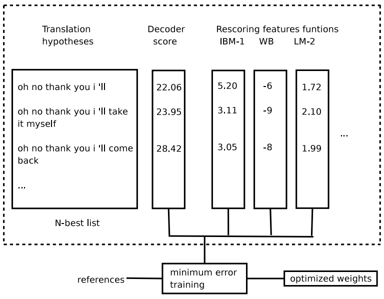

Figure 2: *Standard rescoring framework.*

Multiples translations can be computed by one MT system or by different MT systems. We may assume that different MT systems make different errors due to using different models, generation strategies or tweaks. An investigated technique, inherited from Automatic Speech Recognition, is the so-called system combination that is based on combining the outputs of multiples MT systems.

Given that most experiments of this PhD were done with phrase- and Ngram-based SMT systems, we combined the outputs of both systems using statistical criteria and additional rescoring features.

4. Thesis' goals (hypothesis to be tested)

This PhD is focused on achieving one main objective. Additionally, it addressed complementary challenges without deviating from the original scope.

- **To introduce novel reordering statistical techniques which are able to produce the translation in the correct word order**. Source and target languages may have different order structures. Languages may differ in the basic word order of verbs (V), subjects (S), and objects (O) in declarative clauses. For example, Spanish and English are both SVO languages. German, by contrast, is an SOV language while Classical Arabic and Urdu are VSO languages. Furthermore, there may be other structural differences in word orders between constituents; for instance, modifiers for nouns or verbs may be located in different places. Actually, an important deficit of current SMT systems is the difficult introduction of reordering capabilities. Incorporating them in the search process implies a high computational cost. However, reordering plays an important role, especially in some language pairs, such as Arabic or Chinese to English. The main issue of this thesis consists in developing an efficient reordering technique that can solve statistically the difference in word order of any language pair.

Other complementary objectives are:

- **To propose rescoring and system combination strategies which captures the best quality translations**. Current translation algorithms segment the given source sentence into units and then translate each unit. Therefore, it can become extremely complex to introduce feature functions that deal with information of the entire translated sentence. In those cases, translation may be performed in two steps. In the first step, we compute an N-best list. In the second step, the N-best list is reranked by applying additional features functions. Furthermore, the reranking of hypotheses allows for easier system combination. Different SMT systems approaches lead to different translations. In order to merge N-best lists which have been provided by different systems, we can use feature functions which decide which is the best translation. We contemplate developing and/or introducing feature functions to discriminate translations.

- **To gain efficiency and accuracy in the translation unit vocabulary**. By complementing the reordering objective, the extraction process and reordering techniques must be combined, either at the word or unit levels. Hence, the way the SMT system learns bilingual units plays an important role in translation quality. Here, the main idea is to further study the extraction of bilingual units from parallel corpora taking practical aspects, such as the translation vocabulary sparseness, into account. For some applications with limited memory space (PDAs, mobiles), the number of bilingual units stored in the device should be limited without affecting the quality of translation. In addition, the fewer bilingual units allows for faster translation time. Hence, it is important to look for the best extraction algorithm linking efficiency and quality.
- **To build and maintain a state-of-the-art phrase-based system to compare its performance with the Ngram-based system.** We improved the SMT translation by using the Ngram-based system. To have impact and be relevant to the community, our improvements were demonstrated, when possible, in an in-house phrase-based system that is the most widely used system in SMT.

5. New methods and analysis

This section briefly describes the methods that were proposed to achieve the PhD objectives.

Reordering

This thesis proposes two novel reordering techniques [1]. The first and less complex is briefly described as follows. Given a word alignment, we identify those pairs of consecutive source blocks (sequences of words) whose translation is swapped, i.e. those blocks which, if swapped, generate a correct monotonic translation. Afterwards, we classify these pairs into groups, following recursively a co-occurrence block criterion, in order to infer reorderings. Inside the same group, we allow new internal combination in order to generalize the reorder to unseen pairs of blocks. Then, we identify the pairs of blocks in the source corpora (both training and test) which belong to the same group. We swap them and we use the modified source training corpora to realign and to build the final translation system. We have evaluated this first reordering approach both in alignment and translation quality. In addition, we have used two state-of-the-art SMT systems: a Phrased-based and an Ngram-based. Experiments on the EPPS task show improvements almost over 1 point in BLEU (the standard MT evaluation metric) [2].

The second is the so-called Ngram-based Reordering (NbR) approach which uses the powerful techniques of SMT systems to generate a weighted reordering graph (see the schema in Figure 3). Thus, statistical criteria reordering constraints are supplied to an SMT system, and this allows an extension to the SMT decoding search. NbR allows for a reduction of the vocabulary sparseness of the SMT system during the training phase. The fact of using classes to train the reordering hypothesis (instead of words themselves) allows to generalize in the test phase. Therefore, the NbR technique is able to generate reordering hypotheses of sequences of words which were not seen during training. Additionally, the NbR technique provides a smoothed context-based weight to each reordering hypothesis by taking advantage of the highly developed language model techniques. Although introducing reordering abilities increases the system computational cost, experiments show that using the NbR technique guides the final translation decoding in an efficient manner. Reordering with the NbR technique highly outperforms our monotonic baseline system and a non-monotonic baseline system with a standard distance-based reordering. Improvement in translation performance has been demonstrated with the EPPS task (Spanish and German to English) and the BTEC task (Arabic to English), achieving improvements of 4 point BLEU [3].

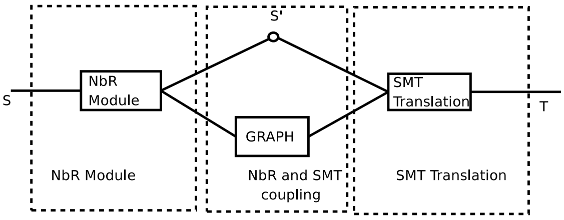

Figure 3: Schema of the NbR and SMT coupling.

Rescoring

The rescoring techniques proposed in this thesis use the continuous space LM which performs probability estimation in a continuous space. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown n-grams can be expected.

The continuous space LM is introduced as a target language model to rescore the n-best lists of a phrase- and Ngram-based statistical machine translation system. We have studied 4-gram language models because we had limited data. With more data, it would be easy to train a continuous space LM with much longer contexts, since the complexity of our approach increases only slightly with the size of the context. Results are provided on the BTEC tasks of the 2006 IWSLT evaluation for the translation direction Chinese, Arabic, Japanese and Italian to English. These tasks provide a very limited amount of resources in comparison to other tasks. Therefore, new techniques must be employed to take the best advantage of limited resources. The results show significant improvement for four different languages pairs and for both systems. The new approach achieves good improvements on the test data; the BLEU score increases by up to 1.9 points.

The continuous space language modeling was successfully extended to smoothing the bilingual language model of an Ngram-based system. The continuous space language model is trained on a bilingual sequence of tuples and it is introduced in the Ngram-based system rescoring. Our method is distinguished by two characteristics: better estimation of the numerous unseen n-grams; and a discriminative estimation of the tuple probabilities. Results are provided on the BTEC task of the 2006 IWSLT evaluation for the translation direction Italian to English. We have chosen the Italian to English task because it is challenging to improve the already good quality of the translation task (over 40 BLEU). Using the neural model for the translation and target language model, an improvement of 1.5 BLEU points on the test data was observed. The described smoothing method was explicitly developed to tackle the data sparseness problem in tasks like the BTEC corpus. Recently, continuous space language modeling applied on the target model (of a phrase-based system) has shown significant improvements when large amounts of data are available (see LIUM site participation in the 2008 NIST evaluation).

System combination

We propose a straightforward system combination method using several well-known feature functions for rescoring the 1-best output of the phrase- and Ngram-based SMT systems, using several n-gram language models, a word bonus and the IBM Model 1 for the whole sentence. The combination seems to obtain clear improvements in BLEU score. We report a structural comparison between the phrase- and Ngram-based system. On the one hand, the Ngram-based system outperforms the phrase-based in terms of search time efficiency by avoiding the overpopulation problem presented in the phrase-based approach. On the other hand, the phrase-based system shows a better performance when decoding under a highly constrained search. We carry out a detailed error analysis in order to better determine the differences in performance of both systems. The Ngram based system produces more accurate translations, but also a larger amount of extra (incorrect) words when compared to the phrase-based translation system. We present another system combination method which consists of concatenating a list of the respective system outputs and rescoring them using the opposite system as a feature function, i.e. the Ngram-based system is used for the phrase-based system and vice-versa. For both systems, including the probability given by the opposite system as a rescoring feature function leads to an improvement of BLEU score [4].

State-of-the-art SMT system

All methods presented were studied to gain efficiency and accuracy in translation and were contrasted with best state of the art systems in multiple MT International Evaluation Campaigns. The work in this events has to be understood in most cases as a team work and it complements the PhD scopes. We particpated in: IWSLT (2005-2008), WMT (2005-2008), NIST (2006 and 2008) and TC-STAR (2005-2007), which means a total of 11 evaluations, presenting the winning system in more than 5 tasks.

6. Project framework

The research presented in this thesis is mainly based on work carried out in several research projects on spoken machine translation: Aliado (2004-2006) funded by the Spanish Government (TIC2002-04447-C02); TC-STAR (2004-2007) funded by the European Union (IST-2002-FP6-506738); Avivavoz (2007-2009) funded by the Spanish Government (TEC2006-13964-C03); and tecnoparla (2007-2009) funded by the Catalan Government.

7. Achievements

Actively pursuing the major thesis objective has lead to the following main research contribution to the SMT field:

- A novel approach for solving the word reordering challenge in an SMT system: a first-pass translation is performed on the source-text, converting it to an intermediate representation, in which source-language words are presented in an order that more closely matches that of the target language. This first translation is performed using an Ngram-based system. Reordering is coupled with translation, which then allows a choice among multiple reordering paths.

Further research contributions are:

- A novel approach for solving local word reorderings in an SMT system. The main limitation is that it addresses reordering in a deterministic way (a fixed reordering is given to the SMT system).
- Experimental work to introduce continuous space language models both in phrase- and Ngram-based SMT systems and its influence in translation.
- A study of two state-of-the-art SMT systems mentioned above. This study leads to system combination at the rescoring level.
- Construction of several SMT systems which were presented at International Evaluation Campaigns. Building a machine translation system is a serious undertaking. The participants are usually provided with a common set of training and test data. Therefore, systems are evaluated under similar conditions, generally with automatic and human measures.

The main techniques and results of this PhD thesis have given place to aproximately 40 international publications [1] including the following publications in JCR journals [3, 2, 4, 1, 5, 6]. The [3] work won the 2009 RTTH BEST PAPER AWARD. Additionally, the PhD itself has been published by VDM Verlag [7]. Finally, a SMT demo system where the PhD author participated very activily is available at *http://www.n-ii.org*. This demo received the 2009 Iberian SLTech BEST DEMO AWARD.

8. References

[1] M. R. Costa-jussà and J. A. R. Fonollosa, “State-of-the-art word reordering approaches in statistical machine translation,” *IEICE Transactions on Information and Systems*, vol. 92, no. 11, pp. 2179–2185, November 2009.

[2] M. R. Costa-jussà, J. A. R. Fonollosa, and E. Monte, “Recursive alignment block classification technique for word reordering in statistical machine translation,” *Language Resources and Evaluation*, p. Accepted for publication, 2010.

[3] M. R. Costa-jussà and J. A. R. Fonollosa, “An ngram-based reordering model,” *Comput. Speech Lang.*, vol. 23, no. 3, pp. 362–375, 2009.

[4] M. R. Costa-jussà and J. A. R. Fonollosa, “Phrase and ngram-based statistical machine translation system combination,” *Applied Artificial Intelligence: An International Journal*, vol. 23, no. 7, pp. 694–711, August 2009.

[5] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, and M. R. Costa-jussà, “N-gram based machine translation,” *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, December 2006.

[6] M. R. Costa-jussà, “New reordering and modeling approaches for statistical machine translation,” *Knowledge and Engineering Review*, p. Accepted for publication, 2010.

[7] ——, *Improving reordering and modeling in statistical machine translation*. ISBN 978-3-639-23568-5 VDM Verlag,Saarbruecken, Germany, 2010.

[1] Please visit the author personal web page at *http://gps-tsc.upc.es/veu/personal/mruiz/*