Using Translation Paraphrases from Trilingual Corpora to Improve Phrase-Based Statistical Machine Translation: A Preliminary Report

Francisco Guzmán Herrera Leonardo Garrido Luna
Instituto Tecnológico de Monterrey
Centro de Sistemas Inteligentes
Monterrey, México
guzmanhe@gmail.com *leonardo.garrido@itesm.mx*

Abstract

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation. Nevertheless, Statistical Machine Translation effectiveness is limited to situations where large amounts of training data are available. Therefore, the broader the coverage of a SMT system is, the better the chances to get a reasonable output are. In this paper we propose a method to improve quality of translations of a phrase-based Machine Translation system by extending phrase-tables with the use of translation paraphrases learned from a third language. Our experiments were done translating from Spanish to English pivoting through French.

1. Introduction

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation [4]. There have been several attempts to improve the performance of such systems. Non-syntactic phrase-based translation systems[9] certainly outperform word-based systems[21]. Nevertheless, Statistical Machine Translation (SMT) effectiveness is limited to situations where large amounts of data are available.

Such a condition, limits the performance of SMT systems over"low density" language pairs [5]. Scarce training data, often leads to a low coverage problem, that is, a low amount of learned translations for a language pair. In this paper we will discuss a method for expanding learned translations by means of a third language, so coverage is augmented and translation quality incremented.

This paper is organized as follows: In Sec. 2, we give an outline of the related work being done in phrase-based SMT. In Sec. 3 we describe the coverage problem and how extending phrase-tables we can tackle this problem. In Sec. 4, we describe thoroughly the translation paraphrases we used in our experiments. In Sec. 5, we explain the methodology followed throughout our experimentation and in Sec. 6 we discuss the results. In Sec. 7, we discuss our results and propose further improvements to our system.

978-0-7695-3124-3/08 $25.00 © 2008 IEEE
DOI 10.1109/MICAI.2007.34

2. Related work

There are several efforts trying to improve translation quality of SMT systems. Many state-of-the-art systems involve the introduction of syntactic information to phrase-based machine translations. For example [12] use depth-one syntactic dependency subtrees using a syntactic parser called SYNTEX. Conversely, [6] make use of alignments built upon linguistic annotations in the form of views, to create several translation models, which they later combine to improve translation results. On the other hand, [1] use Combinatorial Categorial Grammar (CCG) supertags in a factored model phrase-based machine translation. Although they have shown some good results, they haven't detected if their improvements come from supertags or from their reordering models. Similarly, [7] studies the CCG supertags and LTAGs syntactic information effects on phrase-based machine translation. Other approach for including syntactic information was studied by [20]. In they work, they propose a new translation model where they treat alignments as trees that align progressively smaller sentence segments(phrases). They later input this model to a phrase-based decoder and analyze results.
Closely related to the work proposed in this paper, we find [5] who improves translation quality by giving alternatives to broaden coverage of a phrase-based machine translation system through the use of paraphrases. Conversely to our work, they obtain paraphrases by translating Spanish and French to Danish, Finnish, German and other languages; and finding whether a Spanish (or French) phrase has a paraphrase given they have a common translation in other languages. Then they use that information in cases where a Spanish-English or French-English phrase is not found in their phrase-tables. Another difference with this work is that they propose a modification to the log-linear model[15] by including the paraphrase feature.

3. Extending phrase-tables

By increasing the basic unit of translation (from words, or unigrams, to phrases or n-grams), phrase-based translation [9, 21] solved many of the problems of the original word-based systems[4]. For instance, contextual linguistic information (such as concordances and collocations) is memorized to a certain extent. Nevertheless, word dependencies that have never been observed tend not to be handled the right way. Moreover, when words are unknown, there is no satisfying strategy to deal with them, given that translating them unaffected or simply omitting them does not ensure a better quality. Therefore, the broader the coverage of a SMT system is, the better the chances to get a reasonable output are. This corelation between translation quality and data training size is better known as the coverage problem.

The strategy we propose to tackle the coverage problem is to extend phrase-tables that are used for SMT with translation paraphrases learned from a third language. Figure 1 exemplifies this point. For example, when training a system with a Spanish-English bitext, we ended up with *their houses* as the sole translation for the Spanish phrase *sus casas*. On the other hand, when bridging the translation proces through French phrase *leur maison*, we also encountered *their homes* as a possible translation for *sus casas*. Moreover, since *their houses* also appeared when the extension was done, that phrase is reinforced so it is more likely to be translated when encountering Spanish phrase *sus casas*. The same

applies for *situación financiera* which now will have *financial circumstances* as a translation alternative besides the already encountered *financial situation* thanks to bridging through french phrase *situation financière*. Notice that this method of enhancing is bidirectional, that is, it will work when translating from English to Spanish and vice versa.

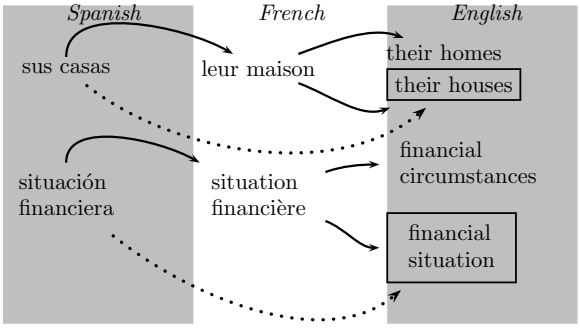

Figure 1. A translation paraphrase example: When translating from Spanish to English with a Spanish-English trained phrase-table, we only get "their homes" English phrase as an alternative to "sus casas" Spanish phrase. However, if using translation paraphrases issued from French, we get "their homes" and "their houses" alternatives

4. Building translation paraphrases

Contrary to meta-phrases which are "literal translations", paraphrases are the expression of information with different words. We could regard paraphrases as different phrases carrying similar meaning. Generation and extraction of paraphrases is a whole field of research within NLP community [17, 3, 2]. Regarding phrase-based SMT, previous work has been done by [5] who use French and Spanish paraphrases extracted from a collection of multilingual bitexts(Spanish-Finnish, Spanish-German, French-Italian, etc) in order to improve coverage and translation quality from Spanish and French to English. Their system works by extending a phrase-based SMT system to include paraphrase probabilities of unknown source language phrases.

4.1. Obtaining translation paraphrases

In our scope, translation paraphrases are the mechanism of preserving meaning through translation. While bridging through a third language, translation paraphrases serve to give more flexible interpretations of source texts, as well as to reinforce translations that are more likely to be good translations regardless of the translation process. An ambitious outlook would suggest that by creating translation paraphrases we are a step closer to an interligual approach, where a universal meaning is carried across languages. Despite such an ideal, our mere intention is to provide a system that increases the output quality of phrase-based SMT between a language pair, and can be an aid in situations where training data is insufficient.

4.2. Building trilingual corpora

In order to train a SMT system, we need what are known as "aligned corpora" or bitexts, which are sets of parallel documents in two different languages; where each line of text in a document written in the first language has a corresponding translation in its parallel document. Although many of the corpora available are aligned with English (Spanish-English, Dutch-English, French-English, etc), there are publicly available tools for creating any language pair corpus by bridging through English. For our research, we wanted to find out how much extra information could be found when training a language triad over the same information. Therefore, we needed to have trilingual aligned corpora, that is, documents aligned in three different languages.

Finding such corpora is not an easy task, given that most of the available parallel corpora are aligned in pairs. Nevertheless, building a trilingual corpora from two corresponding bitexts is rather simple [18]. For that purpose, we used the publicly available Europarl set of aligned corpora [10] and aligned simultaneously the English-French and English-Spanish bitexts to obtain a trilingual English-Spanish-French aligned version of such texts by bridging Spanish-French trough English.

4.3. Shared information of training corpora

When designing a trilingual training set, we need to keep in mind that each language pair belonging to the triad is trained separately. That is, after training an SMT system over a language triad, we end up with three "phrase-tables" (or collections of learned translations), one for each language pair. Therefore it is important to outline how much information it is shared between the different corpora over which we are training each language pair, because the amount of information they share, can later have an effect in translation quality. Therefore we defined a measure of shared information or "information sharing factor" that can be described as the percentage of information common to training the bitexts that share a language, measured in number of lines. For instance, let $C^{l_1}_{l_1 l_2}$ be the l_1 corpus that is used to train language pairs l_1 and l_2 and $C^{l_1}_{l_1 l_3}$ be the l_1 corpus used to train l_1 and l_3. Then the sharing factor between $C_{l_1 l_2}$ and $C_{l_1 l_3}$ is:

$$Sh_{l_1}\left(C_{l_1 l_2}, C_{l_1 l_3}\right) = \left| C^{l_1}_{l_1 l_2} \cap C^{l_1}_{l_1 l_3} \right|. \tag{1}$$

In a more general way, the sharing factor of a language l_i over the set of languages L used to train the system would be:

$$Sh_{l_i} = \left| \bigcap_{l_j \in \{L - l_i\}} C^{l_i}_{l_j l_i} \right|. \tag{2}$$

The mean sharing factor of a system $\bar{Sh}$ would then be:

$$\bar{Sh} = \frac{\sum_{l_i \in L} Sh_{l_i}}{|L|}. \tag{3}$$

Table 1. Example of an extract of a phrase table for an English-Spanish translation. In this table with 2 factors are shown: the phrases in the source language (f), the phrases in the target language (e) and their respective translation probabilities

f	e	$p(f\|e)$	$p(e\|f)$
sus casas	their homes	1.0	0.5
sus casas	their houses	1.0	0.5
sus casas y	their houses and	1.0	1.0
sus casas y sus	their houses and their	1.0	1.0
sus casas y sus hijos	their houses and their children	1.0	1.0

Figure 2 illustrates the concept of the sharing factor between three training sets: Spanish-English, French-Spanish and French-English. In the work discussed in this document our mean sharing factor is set to one $\bar{Sh} = 1$, meaning that our three phrase-tables (en-es, fr-es, en-fr) were obtained from training over the combination of only three corpora ($C^{\text{es}}_{\text{en,es}} \equiv C^{\text{es}}_{\text{fr,es}}$, etc).

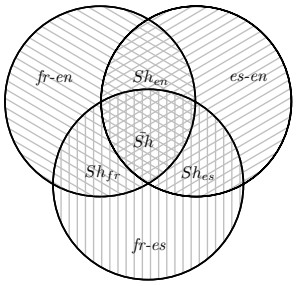

Figure 2. The sharing factor Sh its a measure of how much information it is shared between different sets of training corpora.

4.4. Obtaining translation paraphrases

Having a trilingual aligned training corpora, three phrase-tables were built with the combination of the three languages. In our case, our source language was Spanish, our target language English and our intermediary (bridging or pivoting) language was French. Thus, the phrase-tables we obtained were en-es, en-fr and fr-es. An example of such tables is shown in Tab. 1 .

Contrary to the work done by [5] where paraphrases were extracted from bitexts between Spanish and Dutch, Finnish, Portuguese, etc., our system extracts paraphrases issued from translating phrases from Spanish to French and then from French to English. This is a three step translation. As a result, we ended up with an enhanced Spanish-English phrase-table where new translations for Spanish-English were found thanks to an intermediary translation step through French.

4.5. Calculating probabilities

The probability that certain phrase s of a source language be translated as a target language phrase t is represented by the translation probability $p(t|s)$. Furthermore, the translation probability from the source language phrase to a target language phrase by passing through an intermediary language $p_i(t|s)$can be computed as follows:

$$p_i(t|s) = \sum_i p(t|i)p(i|s). \tag{4}$$

That is, the marginalized probability of translating the source language phrase to an intermediary language phrase and then translating that phrase to the target language phrase. Note that $p_i(t|s)$ represents a translation paraphrase given the intermediary language phrase. In our experiments we computed p_i(en|es) having p(fr|es) and p(en|fr) extracted from our fr-es and en-fr phrase-tables respectively.

4.6. Adding paraphrases to our phrase-table

In order to measure the advantages of using $p_i(t|s)$ as a translation probability, we built a model where the maximum likelihood estimate $p^*(t|s)$ does not rely only in a target/source training but also in intermediaries:

$$p^*(t|s) = \alpha p(t|s) + (1-\alpha)\, p_i(t|s) \tag{5}$$

In a more general way, the model can be extended to include any kind of intermediate operations:

$$p^*(t|s) = \alpha_0 p(t|s) + \alpha_1 p_i(t|s) + \alpha_2 p_{ii}(t|s) + ... \tag{6}$$

Where p_{ii} is a two intermediary step translation and $\sum_i \alpha_i = 1$.

How much extra information do we get from computing $p_i(t|s)$? In our experiments we varied α to observe how translation responds to such variations.

5. Experimental design

In our experiments, we analyzed the behavior of a phrase-based SMT system by adding information obtained through translation paraphrases to enhance phrase-tables for translations from Spanish to English, having French as an intermediary language.

5.1. SMT training

The system we used was based on the log-linear model described in [15]. This model contains eight feature functions, which were tuned using a minimum error rate training (MERT) [14] on a development set to maximize the BLEU score (which is a measure of translation quality) [16].

For extracting phrase translation probabilities we used Giza++ [13] training for each of the corpora pairs. For language model building, we used SRILM toolkit [19] with Kneser-Ney smoothing [8]. Language models were 3-gram, fixed throughout every experiment. We also used Moses decoder [11] to produce translations.

5.2. Training sets

In our preliminary experiments, we created a 10k subset of our trilingual-aligned version of Europarl Corpus [10] of French, Spanish and English documents. This subset was created using the European Parliament documents from January to December 2001 consolidated into one large document for each language. Those documents were afterwards prepared by removing unwanted characters, empty lines and discarding the lines longer than 50 characters. Finally those three documents were truncated to meet our specific corpus size (10k).

Then, we trained the system to obtain three different phrase-tables En-Fr, Fr-Es, En-Es for the set. Note that we used the same set of documents for the phrase-table construction and thus our mean sharing factor was always 1. In our phrase-table training, we considered n-grams up length 7 (that is, phrases with up to 7 words).

5.3. Paraphrase extraction and translation evaluation

Once we had three phrase tables for the training subset, we combined the En-Fr and Fr-Es phrase tables to obtain what we call the paraphrase-table En-Es$_i$. This unification was done by matching French phrases in both tables and computing new probabilities using (4). After that, we needed to consolidate our phrase-table En-Es with our paraphrase-table En-Es$_i$ to get En-Es*. For consolidating those tables, we used (5) and varied α from 0 to 1 in steps of 0.1, to obtain ten different En-Es* phrase-tables. This was done in order to measure how translation quality responded to α changes. Note that (5) has a little tweak. It does not specify what to do when phrases from En-Es do not appear in En-Es$_i$ or vice-versa . We had then the choice of either treating the existing translation probability as the final translation probability, or treating the non existing phrase translation probability as 0 (discounting model). In our experiments we preferred the second approach.

Once we had ten phrase-tables the 10k subset, we did a MERT training [14] over each phrase-table using a 500 line subset of the Europarl development set for Spanish and English to produce optimal configuration files for Moses decoder.

After tuning the decoder for each phrase-table, we translated the Europarl test set using each of the phrase-tables previously obtained. Then, we evaluated the results using the BLEU score [16].

6. Experimental results and discussion

In Tab. 2 we find the results for the series of experiments conduced with the minimum rate error training over the 10k subset. From left to right, we find the overall BLEU score, then the BLEU scores for unigrams, the BLEU scores for 2,3 and 4-grams. As shown, the best performing α is 0.6 with an overall BLEU score of 28.08, which represents only a 3.15% increment in translation quality. Furthermore, individual n-gram maxima can be found at: $\alpha = 0.2$ for unigrams, $\alpha = 0.5$ for bigrams and $\alpha = 0.6$ form 3-grams and 4-grams. As we can see, greater increments are achieved as for longer n-grams, reaching a 5.59% form 4-grams. This may be due to the effect of training with $\bar{Sh} = 1$: we may not be able of obtaining lots of information from unigrams, since they are better detected by En-Es training but, we find better translation paraphrases at higher n-grams.

Table 2. Summarizing table for experiment 1. BLEU score for n-grams against different α.

α	BLEU	B-1	B-2	B-3	B-4
0.0	27.22	57.9	31.8	20.8	14.3
0.1	27.94	58.1	32.5	21.6	14.9
0.2	28.05	**58.4**	32.6	21.7	15.0
0.3	27.87	58.2	32.4	21.5	14.9
0.4	27.81	57.9	32.3	21.5	14.9
0.5	27.97	58.2	**32.5**	21.6	15.0
0.6	**28.08**	58.3	32.5	**21.7**	**15.1**
0.7	27.76	58.0	32.2	21.5	14.8
0.8	27.84	58.2	32.4	21.5	14.8
0.9	27.73	58.1	32.2	21.4	14.8

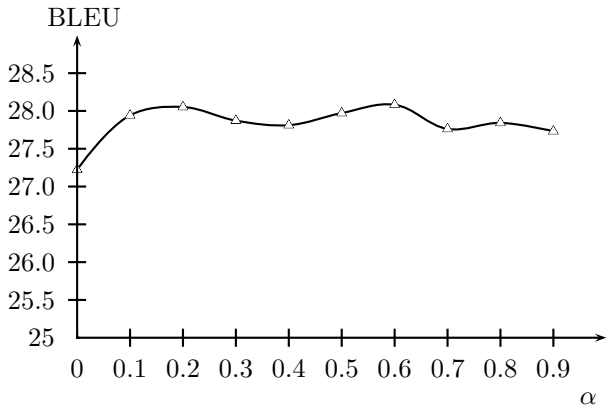

Figure 3. Results of BLEU vs. α

Figure 6 shows that overall BLEU score does not behave in an predictable fashion as we vary α. This may be due to the effect of the MERT training over each of the phrase-tables: each phrase-table is given different factor weights, therefore linear increments in α result in non linear variations in BLEU.

Even if a 10k subset is a rather small collection of data, experiments presented in this work are of an exploratory nature. We are currently working on extending this study to include greater amounts of data. We are confident that better results can be achieved in a matter of time. On the other hand, we are exploring new alternatives to combine the data proceeding from phrase-tables obtained from different training corpora ($\bar{Sh} \neq 1$).

7. Conclusions and future work

In this preliminary study, we have presented a new methodology to ample coverage and improve SMT output quality by the inclusion of information extracted from a third language. Although the results presented in this paper represent small improvements, we have great expectations about translation paraphrases. We are confident that vaster results will be shown when training with totally different corpora, that is, with $\bar{Sh}$ close to zero. For

example, this technology may have an application where resources for training translators for minority languages (such as Mexican Nahuatl) are scarce and limited to alignments with a single majoritarian language (such as Spanish). Therefore, our methodology could be applied to build translators between Nahuatl and English, by exploiting all information gathered through Spanish-English and Nahuatl-Spanish albeit Nahuatl-English information has low availability.

We also believe that there are different directions to explore next. For example: Which bridging language is the best for translating between X and Y? In this experiment we explored a possibility using the Spanish-French-English schema, but Italian or German may work even better as intermediary languages. We do not know yet, but we are confident that in a near future, our following developments will give compelling results that information obtained through a third language is useful.

References

[1] Miles Osborne Alexandra Birch and Philipp Koehn. Ccg supertags in factored statistical machine translation. In *ACL Workshop on Statistical Machine Translation*, 2007.

[2] Regina Barzilay and Lillian Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL 2003: Main Proceedings*, pages 16–23, 2003.

[3] Regina Barzilay and Kathleen Mckeown. Extracting paraphrases from a parallel corpus. In *Meeting of the Association for Computational Linguistics*, pages 50–57, 2001.

[4] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[5] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[6] Jesús Giménez and Lluís Màrquez. Combining linguistic data views for phrase-based SMT. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 145–148, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[7] K. Sima'an H. Hassan and A. Way. Supertagged phrase-based statistical machine translation. In *45th Annual Meeting of the Association for Comp. Linguistics*, 2007.

[8] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184, 9-12 May 1995.

[9] P. Koehn, F. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton, Canada, May 27-June 1 2003.

[10] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. MT Summit 2005, 2005.

[11] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, and Evan Herbst. Alexandra Constantin. Moses: Open source toolkit for statistical machine translation. . In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic,, June 2007.

[12] Philippe Langlais and Fabrizio Gotti. Phrase-based smt with shallow tree-phrases. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 39–46, New York City, June 2006. Association for Computational Linguistics.

[13] F. Och and H. Ney. Statistical machine translation. In *EAMT Workshop*, pages 39–46, Ljubljana, Slovenia,, May 2000.

[14] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7 2003.

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

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics*, pages 311–318, 2002.

[17] Yusuke Shinyama and Satoshi Sekine. Paraphrase acquisition for information extraction. In Kentaro Inui and Ulf Hermjakob, editors, *Proceedings of the Second International Workshop on Paraphrasing*, pages 65–71, 2003.

[18] Michel Simard. Text-translation alignment: Three languages are better than two. In *Proceedings of EMNLP/VLC-99*, College Park, MD., 1999.

[19] A. Stolcke. Srilm – an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver, USA, 2002.

[20] Juan Miguel Vilar and Enrique Vidal. A recursive statistical translation model. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 199–207, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[21] R. Zens and H. Ney. Improvements in phrase-based statistical machine translation. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 257–264, Boston, MA,, May 2004.