Using Translation Paraphrases from Trilingual Corpora to Improve Phrase-Based Statistical Machine Translation

Francisco Guzmán and Leonardo Garrido.
CSI-ITESM

1 Translation Paraphrases

Paraphrases are different phrases carrying similar meaning in one language. 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 good translations; regardless of the translation process.

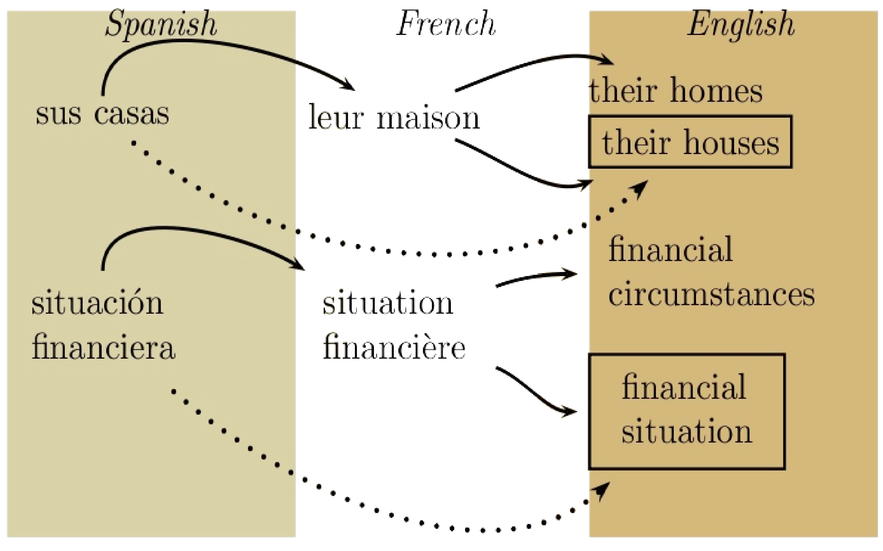

Figure 1. Example of Translation Paraphrases.

2 Trilingual Aligned Corpora

We wanted to measure quality improvements in translation when using translation paraphrases extracted from the same set of documents, in three different languages. Therefore, we needed to have trilingual aligned corpora, that is, every line of text in one document, has a translation in other two documents.

We defined a measure of shared information or "information sharing factor" that can be described as the percentage of information common to training bitexts that share a language, measured in number of lines.

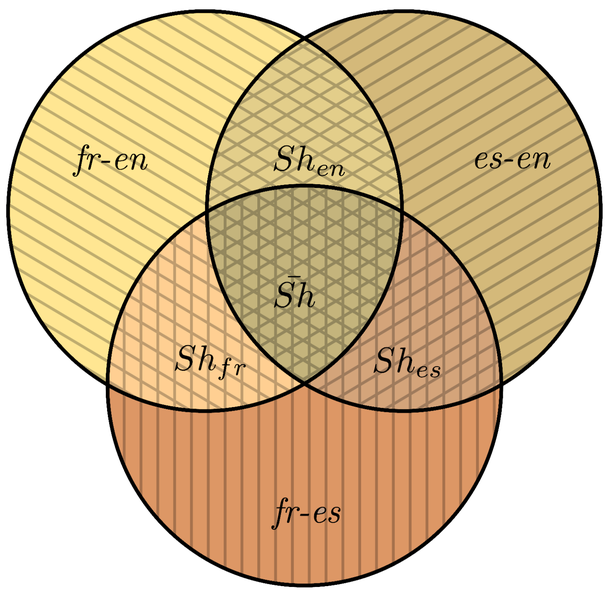

Figure 2. Shared Information between trilingual-aligned corpora.

3 Obtaning Paraphrases

We trained the translation model over the three pairs of languages: En-Es, En-Fr and Fr-Es. Therefore we obtained three phrase-tables.

Our translation paraphrases are obtained from combining the En-Fr and Fr-Es phrase-tables. The translation probability for this combination of phrase-tables is computed as follows:

$$p_{tp}(e|s)=\sum_{f} p_o(e|f)\,p_o(f|s)$$

In order to measure the advantages of using translation paraphrases, we built a model where the maximum likelihood estimate does not rely only in a target/source training but also in translation paraphrases:

$$p_{mix}(t|s)=(1-\alpha)\,p_o(t|s)+\alpha\,p_{tp}(t|s)$$

In our experiments we gradually increased α to obtain several mixed phrase-tables. Finally, we used each of those phrase-tables to translate a set of documents and evaluated their performance using BLEU[1].

4 Results

The best performing system is at α=0.6 with a 3.15% increment in translation quality. Larger quality improvements are achieved for longer n-grams, reaching a 5.59% for 4-grams. Results show better performance for translation paraphrases at higher n-grams.

α	BLEU	B-1	B-2	B-3	B-4
0.0	27.22	57.90	31.80	20.80	14.30
0.1	27.94	58.10	32.50	21.60	14.90
0.2	28.05	**58.40**	32.60	21.70	15.00
0.3	27.87	58.20	32.40	21.50	14.90
0.4	27.81	57.90	32.30	21.50	14.90
0.5	27.97	58.20	32.50	21.60	15.00
0.6	**28.08**	58.30	**32.50**	**21.70**	**15.10**
0.7	27.76	58.00	32.20	21.50	14.80
0.8	27.84	58.20	32.40	21.50	14.80
0.9	27.73	58.10	32.20	21.40	14.80

Table 1. Summarized Results.

5 Conclusions

In this preliminary study, propose a methodology to ample coverage and improve translation quality by using translation paraphrases. Translation paraphrases have applications where resources for training translators are scarce. For instance, our methodology could be applied to build translators between Nahuatl and English, through Spanish.

Acknowledgments

We gratefully thank the support of the Distributed Knowledge and Distributed Agents Research Fund sponsored by ITESM and CONACYT.

References

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