A Machine Translation Model Inspired by Code Generation

Danijel Radošević, Ivan Magdalenić, Darko Andročec, Andrija Bernik, Matija Kaniški
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
{danijel.radosevic, ivan.magdalenic, darko.androcec,
andrija.bernik, matija.kaniski}@foi.hr

Abstract. *In this paper we present our machine translation model that is based on bilingual linguistic phrases and their lexical attributes. Developed model partly arises from our previous work in the field of Generative Programming, and the working prototype was made for Croatian-French and French-Croatian translation. The model is aimed for the general language domain, but it is so far mostly trained for some limited domains like cooking recipes, weather forecast and user manuals for different devices. The proposed model was evaluated according to the matched phrases and by using the BLEU machine translation evaluation method.*

Keywords. machine translation, phrase based translation, BLEU

1 Introduction

In our previous work, we have acquired a lot of experience in a field of Generative Programming (GP) that resulted in SCT generator model (Radošević & Magdalenić, 2011) and its special implementation named Autogenerator (Magdalenić, Radošević & Orehovački, 2013). The idea in the base of SCT was to connect three model elements (Specification, Configuration and Templates), as described in (Radošević & Magdalenić, 2011) to generate the source code of the target program application. Considering that maintaining of configuration rules and code templates represents a kind of background work, remains that the purpose of a generation system is to translate the Specification in the higher level language into a program code (lower level language).

On the other hand, some important experiences in natural language processing were acquired in the development of natural language dictionaries and their use in the education process (Fara & Radošević, 2016). These efforts have resulted in, among others, the Croatian-French and French-Croatian online dictionary[1], with currently 11400 records containing words, their types, pronunciation and examples of use. Within this project, an appropriate search engine for finding translations was developed. Also, the system enables collaboration of more authors in gathering of dictionary terms.

The idea in a base of our machine translation model was to transform the code generation model into a natural language translation model, by using of the developed dictionaries, the search engine and the other findings from our previous research. For this purpose, a prototype[2] of our translation system was made. It can translate in both directions (from Croatian to French and vice versa), but for now it is mostly trained for French to Croatian translation, as well as the conducted tests within this paper. Training and testing in the opposite direction is planned for the future work.

2 Related Work

For machine translation, we can use rule-based, statistical, or hybrid machine translation approaches (Trujillo, 1999). Phrase-based statistical machine translation is the most popular approach in a machine translation research community (Koehn, Och & Marcu, 2003). Koehn et al. (Koehn, Och & Marcu, 2003) have created translation model and decoder to evaluate and compare various translation methods, and their results showed that phrase translation has better performances than word-based methods. Furthermore, Chand (Chand, 2016) has performed an empirical survey of machine translation tools. The tools have tested include rule based systems and statistical based machine translation systems, and it was shown that statistical systems have better performances. Och and Ney (Och & Ney, 2004) have presented the alignment template approach as an extension of a phrase-based machine translation approaches. Their translation approach allows many-to-many relations between words to take the context of words into account.

[1] Croatian-French and French-Croatian online dictionary is available at http://lana.foi.hr/dictionnaire/

[2] Prototype of Croatian-French translator is available at http://lana.foi.hr/traducteur/

Munteanu and Marcu (Munteanu & Marcu, 2005) have proposed a method for discovering parallel sentences in comparable, but non-parallel corpora. They showed that a good-quality machine translation system can be built from scratch by starting with a very small parallel corpus (100,000 words) and exploiting a large non-parallel corpus. Chiang (Chiang, 2005) presented a statistical phrase-based translation model that uses hierarchical phrases (phrases that contain sub-phrases). Cho et al. (Cho et al., 2014) proposed a neural network model called RNN Encoder-Decoder for statistical machine translation. They use the aforementioned model to score each phrase pair in the phrase table, and found that it improves the translation performance in terms of BLEU scores. Och (Och, 2003) analyzed various training criteria that directly optimize quality of statistical machine translation. Maučec and Brest (Maučec & Brest, 2003) offer a comprehensive survey of approaches to coping with Slavic languages in different aspects of statistical machine translation. They claim that languages with a rich morphology pose an especially difficult challenge for machine translation research.

There are some existing tools for statistical machine translations. Apertium (Forcada et al., 2011) is open-source platform for rule-based machine translation. The mentioned platform provides a language-independent machine translation engine, tools to manage the linguistic data necessary to build a machine translation system for a given language pair, and linguistic data for a growing number of language pairs. Moses (Koehn et al., 2007) is an open source toolkit for statistical machine translation that supports linguistically motivated factors, confusion network decoding for the translation of ambiguous input, and efficient data formats for translation models. Morphological, syntactic, or semantic linguistic information are integrated into pre-processing or post-processing steps. Stanford Phrasal (Green, Cer & Manning, 2014) is statistical phrase-based machine translation system written in Java that provides API for implementation of new decoding models, ability to translate phrases with gaps, and also enables the conditional extraction of phrase tables and lexical reordering models.

A similar approach of the phrases-based translation was offered by Zens at al. (Zens, Och & Ney, 2002) as a limited-domain speech translation task from the German to the English language. The authors used bilingual phrases instead of single words in the translation model because the contextual information in single-word based models was excluded for the translation decisions. Some of the today's most commonly used metrics for the evaluation of machine translation are BLEU, ORANGE, METEOR and LEPOR. The basic assumption of all evaluation metrics for machine translation is that the referent translations must be "good" translations and the more a machine translation is like its referent translations then the score is higher. BLEU (Papineni et al., 2002) is a machine translation metric that uses the range from 0 to 1 to evaluate the translation where the referent translation must be provided. Translations will only attain a score of 1 if they are identical to the referent translation. Unlike BLEU the ORANGE (Lin & Och, 2004) machine translation evaluation metrics doesn't require human involvement at all. It uses a set of referent translations automatically, without extra intervention. Generally, a "good" translation should be ranked higher than a "bad" translation based on their scores. Contrary to BLEU, the METEOR (Lavie & Denkowski, 2009) machine translation evaluation metrics uses and emphasizes recall in addition to precision which is in high correlation with human judgments. METEOR also addresses the problem of referent translation variability and features parameterized ingredients. To address some of the weaknesses in the existing machine translation evaluation (non-English to target language translation, low resource language pairs, etc.), Han has proposed the LEPOR metric (Han, 2017). This metric can be easily employed to different language pairs, or new language pairs due to the concise external resources utilization.

3 Machine Translation Model

The idea in a base of our translation model arises from the general code generation model as described in (Czarnecki & Eisenecker, 2000). In such model, the generator is a mechanism that uses some program artifacts like metaprograms and a set of configuration rules to transform the program specification (usually written in some Domain Specific Language, DSL) to produce the program code in a target programming language. In case of the machine translation of natural language, there are some similarities with a code generator, but also some specifics:

- the specification language and the target language are natural languages, both very complex,
- the translation process requires a very big database of translation artifacts (like words, phrases and attributes) and
- configuration rules that lead the translation process are not totally deterministic i.e. depend on probabilities.

Our translation model (Fig. 1.) uses a **bilingual dictionary** in its base. Such dictionary includes words, lexical tags, pronunciation, different meanings and examples of use with phrases. The dictionary is maintained manually, by more editors that mutually collaborate. Each word in a dictionary has one or more meanings in the opposite language, where the order of these meanings determines the ***priority value*** that is used in a translation process to find the most probable translation.

Apart from the bilingual dictionary, there is a **translation phrase database**. This database is also maintained manually, by using of appropriate editor's interface (Fig.2) which is an addition to the translator interface, as shown in Fig. 3. The attributes of the phrases are not entered manually, but updated automatically by updating script. This script uses some attributes inherited from bilingual dictionary (like word type, gender, and number) as well as the attributes extracted from word endings like cases. An example of the phrase record with attributes is as follows:

```
Phrase_HR:prati posuđe Phrase_FR:faire la
vaisselle ATTR: fe:sl fe:inf fd:nf he:gen
```

In the example, the attribute names of Croatian phrases start with '**h**', while the attribute names of French phrases start with '**f**'. The second letter represents the way how the attributes were obtained. Some attributes are inherited from the dictionary (marked by '**d**') and the others were extracted by word endings (marked by '**e**'). Marks after the colon have their meanings as follows: ***sl*** - singular, ***inf*** - infinitive, ***nf*** - feminine noun, ***gen*** - genitive case.

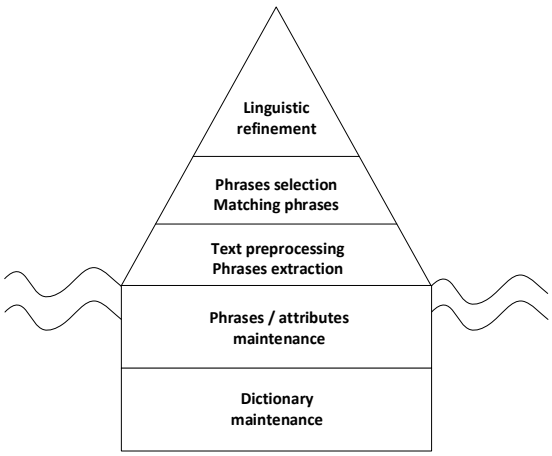

Figure 1. Translation model

The translation process starts by text preprocessing. In this phase, the text for translation is being transformed into a list containing words and separators. Apart to that list, another list, containing attributes of words is also being created, for the purpose of later linguistic refinement.

Phrases extraction is the process of cutting text into phrases of different lengths. The phrases start from each word in the text, containing 1-5 words. **Matching phrases** tries to find the longest phrase with the translation in the translation phrase database. Such phrase is observed as the **candidate for translation**. In some cases, there are more candidates, so a list should be created. The order inside the list depends on the **priority value** of the phrase. If the phrase is inherited from the dictionary, the priority depends on the order of meanings, but if the phrase is entered by translator editor's interface, the priority value is equal to zero (highest priority). Another factor that impacts the order of candidates is the **similarity value** that is used for phrases containing only one word. This value is expressed by the similarity of strings (based on the Levenshtein distance between strings), on a scale 0 - 1, between the candidate for translation and a word from translation phrase database.

Some candidates for translation have to be excluded because of the overlapping problem. In some cases, the phrase can include the beginning of another phrase (but not the whole another phrase). In that case, the general strategy is **shortening the first phrase**, except in a case when the first phrase is in the list of non-shortening phrases. In that case, the second phrase will be excluded.

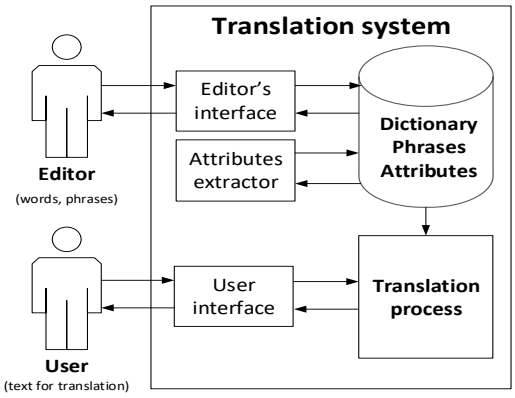

Figure 2. Translation system

After aforementioned phases of translation, there are still some issues to be solved in order to obtain a better translation. It's the **refinement phase**. The refinement is made by changing the priority values of translation candidates. There are, for now, two ways to do that:

- **by increasing the priority of shortened phrases.** In the case of shortening phrases (in order to avoid the overlapping), it's possible to **increase the priority of shortened phrases** according to the similarity between shortened phrases and non-shortened phrases. This enables the usage of phrase context in order to find the most appropriate translation.
- **by using of lexical attributes**. As mentioned before, the phrases have their lexical attributes according to word types, genders, cases etc. Such attributes are also extracted from text to be translated, in a preprocessing phase. Depending on the matching of that attributes with attributes of phrases that are candidates for translation, their priority values can be increased or decreased. This kind of **linguistic refinement** enables translation system to put translations into the right form, according to word type, gender, number or case.

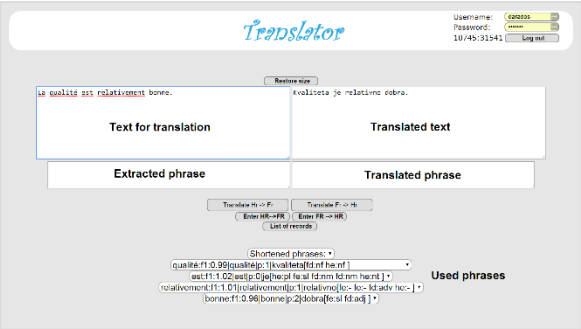

Figure 3. Translation interface

4 Model Evaluation

In order to evaluate our up today's efforts in the building of an effective translator, some tests were provided. There were two groups of tests. The first group was based on the internal measures of the translator, like **coverage of the original text by translation phrases** of different size (containing one or more words). These tests have been carried out on documents that belong to seven different categories to see the differences among them. The second group of testing was carried out using widely accepted **BLEU metrics**[3] (Papineni et al., 2002) (Seljan, Vičić & Brkić, 2012). The purpose was to find the indications of the relationship between internal measures from the first group of testing and the BLEU score, as well as the relationship of the BLEU score achieved by our translator and the widely used Google Translator[4]. The total number of phrases in translator's database was 40000, where 31000 were inherited from the dictionary, and the rest of 9000 were entered by using the editor's interface. The direction of translation process was French to Croatian for all testing.

[3] The tests have been conducted using Tilde Interactive BLEU score evaluator, https://www.letsmt.eu/Bleu.aspx

4.1 Testing using internal measures of translator

There were seven categories of articles (recipes, news, user manuals, fashion and clothes, weather forecasts, computer programming and song lyrics) included in a test, with 10 articles per category, where the size of the documents were around 200 words, with the purpose to find out some information about the phrases used in the translation process. The phrases were categorized into seven types:

F1a - unrecognized word (could be a name or doesn't exist in the translator's database)
F1b - the word exceeds the similarity threshold with the one that exists in the database
F1c - the word is exactly the same as the one that exists in the database
F2, **F3**, **F4**, **F5** - the phrase consisting of 2-5 words were found in the translator's database

The results are shown in Fig.4:

[4] Available at https://translate.google.hr/

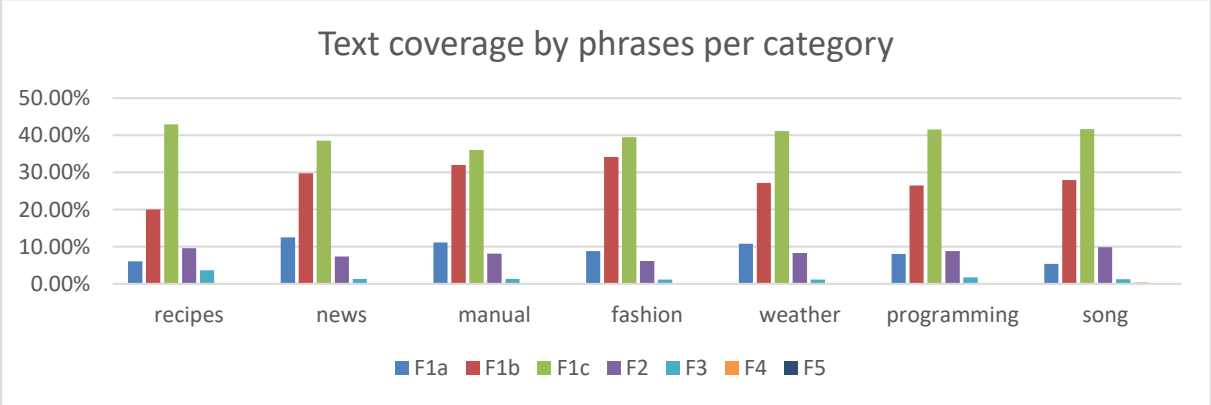

Figure 4. Coverage of text by phrases per category

The results show the high level of words coverage by translator database, because only **5-12%** of words were unrecognized, which includes the names. The most frequent category was ***F1c*** (exact match of words), with **36-43%** of the whole texts, followed by ***F1b*** (words that exceed the similarity threshold), with **20-32%**, and ***F2***, phrases consisted of two words, with **12-20%**. Phrases consisting 3 or more words were relatively rare used (**3-11%** of text for ***F3***, and ***0-1%*** for ***F4*** and ***F5***).

Coverage of text by particular words (***F1a***, ***F1b*** and ***F1c***) and phrases (***F2-F5***) is shown in Fig. 5:

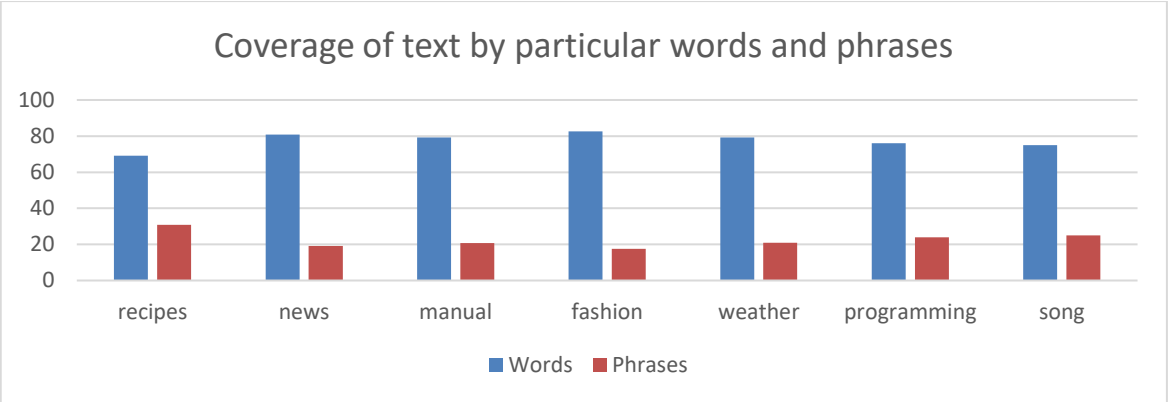

Figure 5. Coverage of text by particular words and phrases per category

The results vary, depending on the category, from **17%** (fashion) to **31%** (recipes). It could be connected with the translator database training, as well as the wideness of vocabulary for different categories.

4.2 Testing by BLEU Score

The BLEU method of machine translation evaluation relies on the similarity between the machine translation and the referent translation (usually human made), giving a score between 0 and 1 (Papineni et al., 2002). Coverage with phrases (***F2-F5***) for the category 'songs' was **25,01%** in the average (Fig.5). Table 1 shows results on a level of particular songs:

Table 1. Coverage by phrases and BLEU score on a level of particular songs

Song	Words	Phrases %	BLEU_test	BLEU_Google
song 01	201	32,84	5,48	27,37
song 02	508	40,16	7,14	15,27
song 03	307	19,54	2,85	7,40
song 04	536	26,12	4,44	10,33
song 05	280	7,86	1,89	10,86
song 06	**252**	**25,79**	**22,74**	**20,46**
song 07	547	17,92	1,47	4,13
song 08	**255**	**24,31**	**10,23**	**6,45**
song 09	355	24,79	11,27	18,17
song 10	181	28,18	3,59	9,60

0,27 Correlation Phrases - BLEU_test
0,48 Correlation BLEU_test - BLEU_Google

It can be seen in Table 1. that the BLEU score of Google Translator was better in 8 of 10 cases. The Pearson correlation between coverage of the translation in phrases and the BLEU score was **0,27**, while the correlation of BLEU scores of the tested translator and Google Translator was higher, **0,48**. These results should be observed as preliminary, because of the very small testing set.

5 Conclusion

This paper presents our machine translation model and prototype of its implementation on the example of croatian-french and french-croatian translation. The model partly arises from our previous work in a field of Generative Programming and our work on building online bilingual dictionaries.

There were some tests performed on the prototype of our translation system. Some tests have used internal metrics of the translation system, like usage of different length phrases, while the others were based on widely used BLEU machine translation evaluation method. The results of the internal metrics show that there are significant differences in the translator's coverage of different categories (e.g. results were much better for cooking recipes than for the other categories). The results of the testing using BLEU method show some positive correlation with the results of internal metrics (0,27), and some positive correlation between results of our translation system and the widely-used Google Translator (0,48).

In our future work, it's planned to include some more elements of our SCT generator model (Radošević & Magdalenić, 2011) into the translation model. For example, using of translation phrases that include some variable parts, like code templates, could significantly reduce the necessary number of such phrases. Some improvements are also possible in the linguistic refinement of translation, including the usage of more phrase context.

Acknowledgments

We would like to thank prof. Višnja Fara and prof. Biljana Grubačević for their work on the online dictionaries and their help in the linguistic part.

References

Chand, S. (2016). Empirical survey of machine translation tools. In IEEE; 2016 [cited 2017 May 29]. p. 181–5. Available from: http://ieeexplore.ieee.org/document/7813653/

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In Association for Computational Linguistics; 2005 [cited 2017 May 29]. p. 263–70. Available from: http://portal.acm.org/citation.cfm?doid=1219840.1219873

Cho, K., Merrienboer, B., Gülçehre C, Bougares F, Schwenk H & Bengio Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. CoRR [Internet];abs/1406.1078. Available from: http://arxiv.org/abs/1406.1078

Czarnecki, K., Eisenecker, U.W. (2000). Generative, Programming: Methods, Tools, and Applications. Addison Wesley

Fara, V. & Radošević, D. (2016). Integrating Technology Into Vocabulary Building. I. International Conference: From Theory to Practice in Language for Specific Purposes (Conference Proceedings), ISSN: 1849-9279, Zagreb, 19.-20. February 2016.

Forcada, M.L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G. & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Mach Transl., p.127–144.

Green, S., Cer, D. & Manning C. (2014). Phrasal: A Toolkit for New Directions in Statistical Machine Translation. In: Association for Computational Linguistics; [cited 2017 May 29]. p. 114–21. Available from: http://aclweb.org/anthology/W14-3311

Han, L. (2017). LEPOR: An Augmented Machine Translation Evaluation Metric. arXiv preprint arXiv:1703.08748. Chicago.

Koehn, P., Hoang H., Birch, A., Callison-burch C., Zens, R., Federico M., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, p. 177-180.

Koehn, P., Och F.J., Marcu D. (2003). Statistical phrase-based translation. Association for Computational Linguistics; p. 48–54. Available from:http://portal.acm.org/citation.cfm?doid=1073445.1073462

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. Machine translation, 23(2), 105-115.

Lin, C. Y., & Och, F. J. (2004). Orange: a method for evaluating automatic evaluation metrics for machine translation. In Proceedings of the 20th international conference on Computational Linguistics (p. 501). Assoc.for Comput Linguistics.

Magdalenić, I., Radošević, D. & Orehovački, T. (2013). Autogenerator: Generation and execution of programming code on demand. Expert Systems with Applications 40.8, 2845-2857.

Maučec, M.S. & Brest, J. (2017). Slavic languages in phrase-based statistical machine translation: a survey. Artif Intell Rev [Internet]. [cited 2017 May 29]; Available from: http://link.springer.com/10.1007/s10462-017-9558-2

Munteanu, D. S. & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, p. 477–504.

Och, F.J. (2003). Minimum error rate training in statistical machine translation. In Association for Computational Linguistics; [cited 2017 May 29]. p. 160–167. Available from: http://portal.acm.org/citation.cfm?doid=1075096.1075117

Och, F.J. & Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation, Computational Linguistics, v.30 n.4, p. 417-449

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318.

Radošević, D. & Magdalenić, I. (2011). Source code generator based on dynamic frames. Journal of Information and Organizational Sciences 35.1: 73-91.

Seljan, S., Vičić, T. & Brkić, M. (2012). BLEU Evaluation of Machine-Translated English-Croatian Legislation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC, pp. 2143–2148. ELRA

Trujillo, A. (1999). Translation engines: techniques for machine translation. London ; New York: Springer; 303 p. (Applied computing)

Zens, R., Och F. J., Ney H. (2002). Phrase-based statistical machine translation. Annual Conference on Artificial Intelligence. Springer Berlin