Term Paper : Phrase-Model and Semantic Roles in Statistical Machine Translation

Keshav Sharma
Department of Computer Science
Birla Institute of Technology and Science (BITS) Pilani
India
f20170140@pilani.bits-pilani.ac.in

***Abstract*—Text Translation is one of the essential aspects of the world of the web. Machine translation is such a vital part of the web as it is time-efficient, relatively cheap, and, most importantly, can translate text across many languages. With the amount of information expanding, the number of internet users increasing, and Artificial Intelligence powering the world, it became necessary for the quality of translation to also improve. This paper presents some techniques used in the process of Machine translation. Furthermore, it also presents how machine translation has improved over the years and also proposes specific ways that can enhance the process of translation.**

***Index Terms*—Statistical Machine Translation, Synchronized Context-Free-Grammar, Phrase-Based Model, Word-To-Phrase Transformation, Phrase Reordering, Bilingual Evaluation Understudy, Semantic Role Labelling, GHKM rule extractions**

I. Introduction

Search engines need to fulfill the queries the users have. Nevertheless, sometimes, document retrieval may not be in our favor. Hence, it is essential to look at the basic building blocks of documents, i.e., sentences. The gap between queries I pose versus the information I need arises from the fact that we, as humans, have an understanding of a structure of sentences along with their meanings. However, this is not an easy task for a machine. Hence previously, machines were not able to perceive ideas or comprehend text written in documents as well as us.

Computer Scientists developed Natural language parsers and analyzers to aid this purpose. These allowed a machine to parse a sentence and give sentences their syntactic structures correctly. These structures are of great use in terms of retrieving documents from the web. Nevertheless, among the top 10 million websites, only 59.3% of them contain the English language. Naturally, there was a desire for Text translation as the user might want to retrieve documents in whatever language he/she wanted. Foundation of Machine Translation (MT) was thus a breakthrough in the field of Computational Linguistics as it paved the way for Text-to-Speech and Language Translation.

The syntactic structures of natural language are of great importance for Machine Translation. It gives a proper structure to the sentences necessary for further processing. However, considering each word on itself and translating it crudely using a dictionary-based Machine Translation was not correct and could alter the sentence in another language. For example, the word *book* might mean reading a *book* as an object or *book* a flight as a verb, which means very different things. Hence Statistical Machine Translation (SMT) was introduced, making use of probabilistic models and capturing the idea of the conditional probability theory. However, the cost of barely working with words was quite high. Hence Phrase-Based SMTs came into the picture, which had many benefits. Considering a phrase, with some reordering could be translated into phrases of different languages. A phrase paints a better picture and gives much more meaning than a word. However, Phrase-based Machine Translation required lots of steps, including converting a treebank in the form of Penn Treebank into some sort of phrasal treebanks, Phrasal reorderings, hypotheses, and Decoding.

With more and more data, the importance of *meaning* became more prevalent. As a result, the study was extended from the Syntactic aspect of a sentence towards the Semantic aspect to improve the quality of text translation. Incorporating semantic features can help model relationships between objects, subjects, and predicates in a much more meaningful way. Not only that, since different languages have different syntactic structures, the positioning of words and their use change in different languages. Hence, a rough translation based on a dictionary or phrasal conversions were not enough. Wu and Fung demonstrated the use of semantic features for this, which included Predicate-Argument pairs and Semantic Role Labelling (SRL). SRL is vital in answering and giving information such as ”who” did ”what” to ”whom” or ”where” and ”how” did ”who” go and do ”what” just by looking at sentences. Hence, SRL would significantly improve the quality of translations as a translation based on just the syntactic structure provides lots of ambiguities and may even change the meaning of the sentence altogether when translated from one language of the other.

Machine Translation techniques and modelings are not only useful for translating texts. Currently, Real-time voice-based machine translation, Neural Machine Translations, and Text-to-Speech conversion use these techniques as their underlying principles. In the upcoming Sections, this paper will attempt to review two fundamental techniques used in Machine Translation i.e., A Tree-To-String Phrase-based model for Statistical Machine Translation and Semantic Roles for String-To-Tree Machine Translation. Further, it will also

highlight specific other directions from past contributors and provide an idea as to where the research might head.

II. Phrase-Based Tree-to-String model for SMT

This paper[1] describes a general method for converting a source Tree into a target String using Phrases and Syntactic transformations by using Synchronous Context-Free Grammars (SCFGs). The Grammar rules in these can be applied to more than one language at the same time and hence are used for capturing grammatical structures of translations. The paper[1] uses the un-lexicalized form of SCFGs to design syntactic transformation models.

Previous researches notable include Tree-to-String approaches. Some of those include Collins et al. (2005)[3], who used handcrafted rules to reorder words, and Nguyen and Shimanzu (2006)[13], who used the concept of Probabilistic CFGs (PCFGs). Apart from these, Liu et al. (2006, 2007)[9][10] used templates on Tree-to-String alignments instead of SCFGs, which was later made better by proposing a Forest-to-String model. It is to note that all these researches used source language syntax and parser. String-to-Tree approach by Yamada and Knight (2001)[17] and Galley et al. (2006)[6] using target language's treebanks and parser also give insight about translating using reordering and specified rules.

This paper[1] focuses on a phrasal method of translating text. Some fundamental properties mentioned are syntactic transformation, which also uses a word-to-phrase transformation model, and a phrase reordering model. Two new stages, a Word-to-Phrase Tree transformation model and a Phrase Reordering Model, are added to the existing Phrase-based SMT architecture. Source tree gets converted to a Source-phrase tree, followed by a reordering phase and finally converted to the target phrase tree. The paper[1] briefly describes the two essential additions i.e., Word-to-Phrase Tree Transformation and Phrase-Reordering Model.

A. Word-to-Phrase Tree Transformation

In the formalized Penn Treebank Tree Structure, CFG rules consist of either a sequence of non-terminals or a single terminal (a single word) on the right-hand side of a Grammar rule. However, for Phrase-based Models, the right-hand side must be a sequence of words satisfying specific properties. The algorithm for transformation makes use of a head symbol on the right-hand side. Part Of Speech (POS) tags play an essential role in determining the head node. After making the head nodes and word spans of each node, replace several subtrees by phrasal nodes. Drop the unwanted nodes, and then generate a phrase-tree for each sentence. The algorithm used produces Unique, connected, and flat trees in a deterministic way, ensuring no ambiguities. The flatness of the tree helps in minimal traversal. The paper[1] also mentions specific dependency rules. These transformations are carried out probabilistically using statistical scoring mechanisms and conditional probabilities of source-parsed and word-aligned bitext and thus prevent from generating ambiguous and erroneous phrase trees.

B. Phrase Reordering

The paper[1] incorporates SCFG rules for reordering of phrases and also makes use of un-lexicalized rules. Learning algorithms are derived from the works of Nguyen and Shimazu (2006)[13]. The paper[1] also uses various tools such as Word alignments, broad coverage parsers, and training data to develop this model. This model is even applicable to those language pairs wherein the target language is poor in resources.

C. Decoding Phase

The paper[1] suggests using a log-linear transformation model for encoding possibilities of a source phrase getting translated to the target language's phrase with specific parameters as described. The paper[1] suggests collecting additional Translation options before this phase for a faster look-up and memory-efficient phrase translation table. Hypothesis testing is performed on each of these translation options. The decoding algorithm is run on this. This algorithm is divided into three steps that occur recursively. Step one distributes translation options to all nodes. After checking valid sequences, the translation of nodes occurs in a bottom-up manner. Syntactic transformations occur, followed by reordering of phrases. Finally, phrases are translated using a monotonic decoding procedure.

D. Experimental Setup and Results

The authors[1] use Conversation, an English-Vietnamese corpus, and Reuters, an English-Japanese corpus for training. Various tools are used for training purposes, as mentioned in the paper[1]. Nearly 17500 and 34500 rules were used for Conversation and Reuters, respectively. These numbers are minimal as compared to the previous methods because many of the CFG rules did not need reordering rules. The paper[1] also suggests solving an issue for the near future i.e., Markovization technique for SCFGs, which has been further researched by Micheal and Kathleen (2007)[11] on Sentence Compression systems.

The Bilingual Evaluation Understudy (BLEU) score is employed as a metric to check the improvement. BLEU scores are a significant parameter to judge the proximity between the translated output by the machine to that of a knowledgeable human translator. After evaluation, normalizing operations are performed to get a final score. Syntax-Directed full-parsing Method, as described, is compared to a phrase-based system already in use. On both the corpora, SD methods outperform that of a phrase-based system with high statistical significance. The paper[1] compares BLEU scores with and without Word-To-Phrase transformation showing a decrease of about 0.5 on both the corpora. The paper[1] then experiments with the level of Syntactic Analysis performed and then looks at the translation quality to find out that it increases with the increase in syntactic level. Since the methods used in the paper[2]

provide much faster methods for Tree-to-String systems, the paper[1] has recommended its use in areas like web translation. The paper[1] also shows that any Phrase-based SMT generally outperforms baseline systems.

The paper[1] concludes by giving some insight into related works that can be used as translation units in this approach, such as n-gram by Jos et al.(2006)[7], factored phrasal translation by Koehn and Hoang(2007)[8] or treelet by Quirk et al.(2005)[14]. The paper[1] recommends it as a possible future study and also recommends the use of n-best trees as input for handling adjunct attachments in the given phrases.

III. Semantic Roles for String-to-Tree Machine translation

The previous section dealt with the process of converting the syntactic tree into a phrasal tree to translate phrases into the target language. However, it does not use the semantics of the sentence. Harnessing only the syntactic properties is not enough to translate today's information. It is crucial to model relationships between words occurring and how they affect each other's presence. Another issue that was not addressed was that since Tree-to-String Translation algorithms use parse trees as their base, the structure of trees often constraint them. On the other hand, a String-to-Tree system generally can be modeled in many different ways because of the free-structure of a string, which allows much more scope towards this area. In the second paper[2], we review how to incorporate semantic features during the translation process. Notably, we see how a string in the source language is converted to a treebank in the target language along with semantic predicate-argument features.

In the past, Liu and Glidea (2010)[5] introduced two types of semantic features for tree-to-string MT. These involved reordering and deletion of semantic roles. Xiong et al. (2012) integrated semantic predicate-argument features of the source language into a phrase-based SMT. Wu et al.(2010)[16] used a different Head-Driven Phrase Structure Grammar (HPSG) parser, adding semantic representations to their translation rules. This paper[2] uses semantic role labels to strengthen the string-to-tree translation system. The approach described in the paper[2] helps improve the BLEU score. It uses GHKM style translation rules where the target has been parsed and labeled with semantic roles. Algorithms help to match the string from the source language correctly with proper ordering and semantic matching to the target parse tree for the training data.

A. Understanding Semantic Roles

Semantic Role Labelling (SRL) is the process of identifying all the arguments in the sentence which relate to a given predicate and label them based on their relationships. SRL helps in various aspects such as Question Answering, Information Retrieval and Machine Translation. If a machine can adequately identify the meaning and semantics of a query, it will lead to better search results. For example, let us consider the sentences "The robot broke my favorite mug" and "My mug broke into pieces". In the first sentence, *mug* appears as the object, whereas in the second sentence, *mug* appears as the subject. Syntactically *mug* appears to be different in both these sentences. However, on asking the question "Which thing was broken", *mug* is the answer in both the cases. Hence, both *mug*s are semantically the same.

We can assign semantic roles to each of the words based on how they relate to a verb, say *break*. For example, the arguments can be the breaker, the thing broken, the instrument used to break, the final state of the thing. Glidea and Jufrafsky (2000)[4], Srikumar and Roth (2011)[15] and many others made significant contributions in the field of automatic semantic labeling. Semantic Role labeling was considered a hard problem. Some challenges it posed were Syntactic alternation, Prepositional role attachment, and Long-range dependencies. However, according to the paper[2], recent advancements have made semantic label predictors more than 90% accurate. There are two methods described in the paper[2] which incorporate semantic labels into a String-to-Tree SMT system.

B. Using Semantically Enriched GHKM rules

The target Trees are tagged with semantic roles in the training corpus, and then translation rules are extracted. Semantic labels are attached to the root of the sub-tree that it is labeling. After this, standard GHKM rule extractors are used in addition to the new semantically enriched rules to translate.

C. Complete Semantic Rules with Added Feature

Special set of translation rules are extracted using Semantic Roles forming small tree fragments on the target side. These rules model a complete semantic structure for each predicate. Decoders and GHKM rules use these structures in the later stages. The GHKM component of the system is modified to extract semantic rules for each predicate. Later, all roles are combined, and a single semantic rule is applied to each predicate discarding unnecessary ones. These rules are then matched onto the target tree side using one-to-one correspondence with labels and predicate. In the end, original GHKM rules are used, and the source string is converted to the target parse tree.

D. Experimental Setup and Results

The paper[2] uses a Chinese-English parallel corpus. PropBank, a research project, added predicate-argument relations to the syntactic trees of the Penn Treebank. It consists of some Core roles, which are verb-specific and are defined in frame files. In other words, different verbs have different argument relationship and play different roles. These also include Adjunct roles describing modifier attributes for each verb. The paper[2] uses PropBank for semantic role labeling with a precision of 90% and a recall of 88%. The paper[2] describes the experiment done using three methods, first being the baseline method (Standard SMT) and the second and third using the Methods described above. The number of rules

used for translation and decoding naturally increased and used nearly 1.42 million translation rules.

The results were tabulated and for the third Method (Complete Semantic Rules with Added Feature), new features were given random weights for various evaluations. The paper[2] then evaluates BLEU scores of the translation. An example comparing translation from the baseline method and the Method is described below.

Sentence : in the new situation of the millennium, the development of Asia is facing new opportunities.
Baseline : facing new opportunities in the new situation in the new century, the development of Asia.
Method 3 : under the new situation in the new century, the development of Asia are facing a new opportunity.

The paper[2] gave the best BLEU score achieved as 25.92. It also states that one reason why Semantically enriched GHKM rules did not give out proper results was that the Semantic role labeling of the corpus would have disconnected the roles from predicates across sentences. Thus the semantic roles would not be correctly connected to the proper predicate during decoding. The paper[2] signifies how important it is to use semantic role labeling as the difference in the baseline scores and the Method adopted to show a significant improvement statistically. There are some cases where the baseline generates wrong order, whereas the third Method works as pointed out by the paper[2]. In conclusion, the paper[2] shows the methods used and draws comparisons between existing methods and that of the paper under consideration [2], showing complete predicate-argument structures can improve the quality of machine translation. It was important for the paper[2] to consider long sentences for this purpose as BLEU scores tend to be biassed towards shorter text. However, the paper[2] considers MERT for optimizing the number of parameters. For future research purposes, these ideas can be implemented by better algorithms such as PRO for ranking purposes.

There is, however, some future research that can improve the SRL model. A long-term plan for improving SRL can be a Question-Answer Driven SRL (QA-SRL) on the training data, which is a simple annotation scheme, wherein given a sentence and a verb, questions related to the verbs are asked and answered until all Q/A pairs are exhausted. This can lead to a better understanding of predicate-argument pairs and machines will genuinely understand the meaning of a query. Evidence suggests that this approach will improve the Question Answering software and ChatBots. Compared to PropBank, the Q/A structure corresponds nicely to the predicate-argument structure. The answers correspond to the arguments in SRL, and the questions are similar to semantic Roles. Since we are using natural human language questions and answers, we would not need a predefined set of roles. Speech-to-Text Translation, ChatBots, and Text-Translation would significantly improve if we perform the Semantic Role labeling using QA-SRL. This, in addition to the approaches used yet, will provide better results as it brings Machine Translation to a new direction projecting the human curiosity.

REFERENCES

[1] Thai Phuong Nguyen, Akira Shimazu, Tu Bao Ho, Minh Le Nguyen, and Vinh Van Nguyen. 2008. *A tree-to-string phrase-based model for statistical machine translation*. In Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008), Manchester, England, August. Coling 2008 Organizing Committee.

[2] Marzieh Bazrafshan and Daniel Gildea. 2013. Semantic roles for string to tree machine translation. In *Proceedings of ACL.*

[3] Collins, Michael, Philipp Koehn, and Ivona Kucerov˘a. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, MI.

[4] Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of ACL- 00*, Hong Kong, October.

[5] Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *COLING-10*, Beijing.

[6] Galley, M., Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, Ignacio Thayer 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proceedings of ACL.*

[7] Jose B Marino, Rafael E Banchs, Josep M Crego, Adri´a‘ De Gispert, Patrik Lambert, Jose AR Fonollosa, and Marta R Costa-Jussa. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

[8] Koehn, P. and Hoang, H. 2007. Factored translation models. In Proceedings of EMNLP-CoNLL. 868–876.

[9] Liu, Y., Qun Liu, Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. In *Proceedings of ACL.*

[10] Liu, Y., Yun Huang, Qun Liu, and Shouxun Lin 2007. Forest-to-String Statistical Translation Rules. In *Proceedings of ACL.*

[11] M. Galley and K. McKeown, "Lexicalized Markov grammars for sentence compression," in *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 180–187, 2007.

[12] Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. Computational Linguistics, 34(2):145–159, June.

[13] Nguyen, Thai Phuong and Akira Shimazu. 2006. Improving Phrase-Based Statistical Machine Translation with Morphosyntactic Transformation. *Machine Translation*, 20(3): 147–166.

[14] Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, MI.

[15] V. Srikumar and D. Roth. 2011. A joint model for extended semantic role labeling. In *EMNLP*, Edinburgh, Scotland.

[16] Xianchao Wu, Takuya Matsuzaki, and Jun'ichi Tsujii. 2010. Fine-grained tree-to-string translation rule extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA. Association for Computational Linguistics.

[17] Yamada, K. and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL.*