View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Waseda University Repository

Learning Translation Rules for a Bidirectional English-Filipino Machine Translator

Michelle Wendy Tan, Bryan Anthony Hong, Danniel Liwanag Alcantara,
Amiel Perez, and Lawrence Tan

[1] 2401 Taft Avenue, Manila, Philippines 1004
{mwendygt, bashx5, dlalcantara_bscs, rebirthproject, zymeth02}@yahoo.com

Abstract. Filipino is a changing language that poses several challenges. Our goal is to develop a bidirectional English-Filipino Machine Translation (MT) system using a hybrid approach to learn rules from examples. The first phase was an English to Filipino MT system that required several language resources. The problem lies on its dependency over the annotated grammar which is currently unavailable for Filipino, which makes reverse translation impossible. Phase 2 addresses this limitation by using information taken from English and Filipino POS Taggers. The seed rules are generated by aligning the POS tags from English and Filipino examples, including their constraints. To perform compositionality, the system deduces the constituent labels by using the longest adjacent POS tags found in both the English and Filipino rule. The system groups together similar rules and generalizes it to encompass a wider range of unseen examples.

Keywords: Rule Generation, Compositionality, Rule Generalization, Bilingual Corpus, Example-based, Rule-based, Machine Learning, Machine Translation, Natural Language Processing

1 Introduction

Filipino is the national language of the Philippines. The 1987 Constitution of the Philippines stated that "The national language of the Philippines is Filipino. As it evolves, it shall be further developed and enriched on the basis of existing Philippine and other languages." This characteristic poses problems for developers of Filipino machine translators.

Our goal is to develop a bidirectional English Filipino machine translation system. Our approach is to use a hybrid of rule-based and example based paradigms by learning rules from examples. Several researches have incorporated machine learning such as those by [1], [2], which automatically generates these rules from bilingual corpus. This technique is most applicable for languages with limited resources such as Filipino.

The first phase of development was an attempt to build an English to Filipino Machine Translator based on the Seeded Version Space Learning presented by [1]. Using the same approach, several issues were realized for the development of the second phase, the Filipino to English translation.

2 Phase 1: English to Filipino Translator

Phase 1 focused on the development of an English to Filipino Translator. It combines rule-based and example-based paradigm. Since the goal was to develop a bidirectional MT system, the hope was to use the same architecture for the development of the reverse translation.

2.1 Phase 1 Learning Module Architecture

Phase 1 has two modules. The training module learns rules based on examples found in the bilingual corpus. In turn, the translation module uses the rules to produce the Filipino sentence based on the English input. Refer to [3] for the system's architecture for the learning phase.

The learning module begins by reading each sentence pair in the bilingual corpus through the Sentence Tokenizer. Only the English sentence is passed to the Lexical and Morphological Analyzer. These are passed to the English Parser where it produces all possible parse trees of the English Sentence.

The seed rule generator builds flat rules based on the English parse tree. Using the lexical alignment, the Filipino rules are established. The Filipino rule adapts the constraints found in the English rule.

Each generated seed rule is then analyzed for possible compositionality. It determines which low-level rules are able to produce higher level representations. This makes up the constituent phrases in the rule. Finally, generalization is performed by merging similar rule constraints. The algorithm is based on the Seeded Version Space Learning algorithm presented in [1].

2.2 Phase 1 Results and Issues

The training phase was tested by an English-Filipino linguist who evaluated sentence pairs that generated over 140 rules. From this, the linguist evaluated that 6% of the extracted rules as incorrect due to incorrect Filipino translations in the corpus, 20% were incorrect because the sentences were not accepted by the parser, while 74% were rated as correct and accurate extraction of rules. A human English-Filipino translator also tested the effects of learning using Subjective Sentence Error Rate. A significant increase in the quality of translation was observed when additional sentences were fed into the learning module

Although the results are promising, Phase 1 is only applicable for sentences accepted by the parser containing the given CFG based on [4]. Currently, the CFG accepts only imperative and declarative sentences. If the parser does not accept the sentence, the system translates it on a word-per-word basis. Although the bilingual corpus is easy to obtain, because of the restrictions of the parser, the amount of useful sentences in the bilingual corpus diminishes. As such, learning is not maximized since the CFG is restricted.

3 Phase 2: English Filipino Bilingual Translator

The original intent was to use the same architecture for Phase 1 in the development of the Filipino to English MT system (Phase 2). Although several problems were identified with Phase 1's approach, a single issue was found to be irresolvable to continue Phase 2. This is the requirement of a CFG with annotations for Filipino. Since the available CFG for Filipino is limited, Phase 2 had to revise the architecture to remove the need for the CFG.

3.1 Phase 2 Learning Module Architecture

Phase 2's approach is to develop a single training module that learns both English to Filipino and Filipino to English transfer rules. Instead of using a parser, Phase 2 makes use of POS Taggers (refer to Figure 1). Note that the English and Filipino POS Taggers are used for learning both English-Filipino and Filipino-English transfer rules.

Phase 2 incorporates a semantic analyzer based on semantic classes found in the lexicon. Given the tagged English and Filipino sentences, the analyzer determines the semantic classes of each word which will be used during the Seed Rule Generation Phase. Since there is no English parse tree, the seed rule generator, compositionality, and generalization modules are different from that of Phase 1.

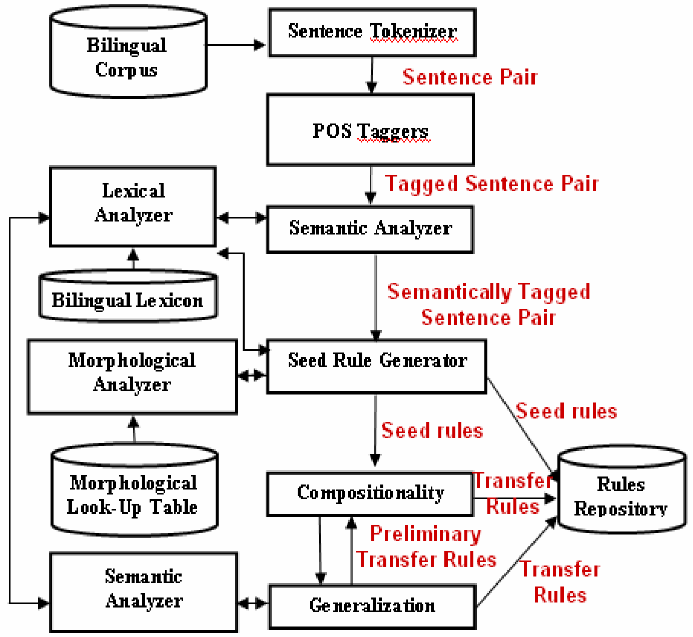

Fig. 1. Phase 2 Learning Phase Architecture.

The first step in learning is alignment. After passing the sentence pair into its corresponding POS Tagger, the next step is to align the POS tags. Phase 2 uses a commercially available English Tagger and a locally developed Tagalog Tagger called TPOST [5]. The alignment is performed by taking the sentence pair and mapping each word in the source language to its corresponding translation in target language. The process is highly dependent on the lexicon and morphological analyzer.

Seed rules are built from each translation pair successfully aligned. Seed rules define the token sequence, expressed as a combination of POS tags and possibly constant words, token constraints and alignment scheme of a translation pair which have been previously aligned. The next phase is compositionality. The goal is to infer rules of higher syntactic structure. The system does this by grouping together POS labels, and forming them together as higher constituent labels. The labels are generated automatically. The basic idea is that by forming compositional rules, a wider variety of examples can be covered. The focus of generalization is to provide the set of the transfer rule a greater range of transfer possibilities than what can be found from just within the corpus. The system does this by identifying similarly co-occurring labels between generalization candidates, and forming them together as generalized labels.

3.2 Phase 2 Results and Issues

A translation module was developed for the testing of the Phase 2. The 10–fold cross validation evaluation method was used [6] on approximately 1500 sentence pairs. The values were tweaked 90% training – 10% testing, 80% training – 20% testing, and 70% training – 30% testing. Three automatic evaluation methods are used: BLEU [7], METEOR [8], NIST [9]. The results are shown in Table 1.

Based on the results, the 80%-20% and 70%-30% averages are almost equal. The 90%-10% average is higher because of a larger training set. It can be observed that English to Filipino translations had a higher score than its Filipino to English counterparts. Filipino to English translations did not fare so well because of the Filipino Tagger. That would mean that a learned functional rule in Filipino will have a lesser scope as compared to a functional rule in English. Therefore, when translating Filipino sentences, there will be a lesser chance that a functional rule will be applied.

Table 1. Summary of evaluation results. Averages are shown for each training-testing ratio.

Training% -Testing%	BLEU - Eng to Fil	BLEU - Fil to Eng	METEOR - Eng to Fil	METEOR - Fil to Eng	NIST - Eng to Fil	NIST - Fil to Eng
90%-10%	0.0573	0.0520	0.2740	0.2308	1.5310	1.4360
80%-20%	0.0463	0.0401	0.2529	0.1809	1.4507	1.2355
70%-30%	0.0481	0.0430	0.2601	0.1832	1.4705	1.2358

4 Conclusion

Filipino is a diverse and changing language. The project focused on the development of a bidirectional English Filipino MT system. During the first phase however the architecture's dependency over the CFG posed problems with the next phase of development which required the reverse translation of Filipino to English A completed Filipino CFG with annotations is currently unavailable making the work around posed by the Phase 2 to be more practical for resource-limited language such as Filipino.

Using commercially available POS Taggers is a more acceptable system requirement. Furthermore, the second phase incorporated semantics through the definition of semantic classes in the lexicon.

The approach of phase 2 still faced with challenges in terms of the lexicon. It is still a struggle to develop a completed lexicon with information that is needed by the system. Semantic classes are taken from WordNet 2.1 [10] by manually entering English and Filipino equivalent semantic classes into the lexicon. Aside from this, co-occurring words generated by an automatic lexicon extractor [11] is needed for semantic analysis. It is then recommended that future research focus on works such as extracting semantic classes for Filipino language and morphological analyzer for Filipino

References

1. Probst, K.: Semi-Automatic Learning of Transfer Rules for Machine Translation of Low-Density Languages. In Proceedings of the 7th ESSLLI student session at the European Summer School in Logic, Language, and Information (ESSLLI-02) (2002)
2. Probst, K.: Automatic Learning of Syntactic Transfer Rules for Machine Translation. Retrieved December 18, 2005, from http://www.cgi.sc.cmu.edu/ People/kathrin/Research/ SummaryOfProposal.pdf (2003)
3. Ang, R. J., Bautista, N. G., Cai, Y. R., & Tanlo, B. G.: Translation With Rule-Learning. Philippines: Undergraduate Thesis, De La Salle University Manila (2005)
4. Borra, A., & Roxas, R.: IsaWika: A Machine Translation from English to Filipino, A Prototype. University of the Philippines (1997)
5. Rabo, V.: TPOST: A Template-based, N-gram Part-of –Speech TAGGER for Tagalog. MS Thesis. De La Salle University-Manila (2004)
6. LIS - Rudjer Boskovic Institute: Evaluation of Models. Retrieved July 2006, from http://dms.irb.hr/tutorial/tut_mod_eval_4.php (2002)
7. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J.: Bleu: a Method for Automatic Evaluation of Machine Translation. Retrieved June 2005, from http://acl.ldc.upenn.edu/P/ P02/P02-1040.pdf (2001)
8. Banerjee, S., & Lavie, A.: METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. Retrieved May 2006, from http://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf (2005)
9. Culy, C., & Riehemann, S.: The Limits of N-Gram Translation Evaluation Metrics. Retrieved May 2006, from http://www.amtaweb.org/summit/MTSummit/FinalPapers/79-Culy-final.pdf (2001)
10. Miller, G. A., Fellbaum, C., Haskell, B., Langone, H., Tengi, R., Wakefield, P., & Wolff, S.: WordNet: a Lexical Database for the English Language. Retrieved December 15, 2005, from http://wordnet.princeton.edu
11. Lat, J., Ng, S., Sze, K., & Yu, G.: AEFLEX: Automatic English Filipino Lexicon Extractor. Undergraduate Thesis. De La Salle University-Manila (2006)