

Automatic Correction of French to English Relative Pronoun Translations using Natural Language Processing and Machine Learning Techniques

Christina George and Nathalie Japkowicz
School of Information Technology and Engineering
University of Ottawa
cgeorge@site.uottawa.ca / nat@site.uottawa.ca

1. Introduction

Machine translation is an area of computational linguistics in which researchers have sought to make significant improvements. In particular, many researchers have focused their interest on statistical machine translation (SMT) models to obtain accurate translations from source languages to English. The use of SMT models may be attributed in large part to Brown et al. (1990), who developed an approach for the translation of single sentences using language and translation models. Since that time, the SMT approach has been modified and expanded by several researchers, including Marcu and Wong (2002) who developed a phrase-based translation model, and Yamada and Knight (2001) who proposed a syntax-based SMT model to include structural and syntactic aspects of language into the translation model.

Although phrase-based and syntax-based SMT approaches have proven to be effective, one important aspect of many of these methods is that they are global – that is, they encompass the entire translation process. With this in mind, one of the main goals of this research is to examine local methods that take the output of SMT systems and attempt to improve on it, specifically with regard to the translation of sentences containing relative pronouns and relative clauses. The mistranslation of relative pronouns is often due to incorrect anaphora resolution, or failure of the system to recognize differences in sentence structure between the source and target languages. The resulting translation is often confusing for the reader and warrants further investigation.

2. Categories of Relative Pronoun Errors

To determine the types of errors that occur with French to English relative pronoun

translations, a preliminary dataset of 72 translated sentences was collected using “Online-Translator”, a machine translation system available through the World Wide Web. Although this system is not a statistical one, it was selected because the English translations were more comprehensible than those produced by Babel Fish in 43 out of 72 instances. It should be noted that each of the sentences in the data set contained at least one incorrectly translated relative pronoun and possibly other types of errors. The translation errors were broken down into several categories as shown in Table 1, with the incorrect relative pronoun produced by the machine translation system and the correct relative pronoun as determined by the authors.

| Incorrect Pronoun from Machine Translation System | Correct Pronoun from Human Analysis | Number of Occurrences |
|---|-------------------------------------|-----------------------|
| Whom | Which | 4 |
| Who | Which | 6 |
| Which | Who | 36 |
| Which | That | 3 |
| Which | Whose | 2 |
| What | Which | 15 |
| What | That | 1 |
| That | What | 1 |
| That | Than | 2 |
| That | Who | 1 |
| Than | That | 1 |
| Total | | 72 |

Table 1 - Breakdown of Relative Pronoun Translation Errors

3. A “Generalizable” Approach using Machine Learning & Natural Language Processing

The primary goal of this research is to correct obvious problems related to relative pronoun machine translation using a process that combines Natural Language Processing (NLP) and Machine Learning (ML). NLP will be used to

study the syntactic and semantic structure of the translated sentences. More specifically, the clauses preceding and following the relative pronoun will be examined in order to select various semantic and syntactic descriptors of entities in the sentence. Once these attributes have been selected, the sentences will be encoded in a vector containing the value for each attribute as well as for the class, which is the correct relative pronoun for the sentence as determined by the authors in Section 2. ML algorithms will then be used to learn the system and to determine whether these features are useful in correctly classifying the sentences according to the appropriate class.

4. Attribute Selection

The selection of attributes for the automatic correction of relative pronoun translations was developed using Advait Siddharthan's (2002) method for resolving relative clause attachment ambiguities with Word Net and machine learning techniques as a starting point.

Siddharthan defines a vector of thirty-four binary features for each sentence containing one of the pronouns *who* or *which* and that is preceded by the phrase structure **NP1 Prep NP2**. These features are based on the Word Net classes of the noun phrases and form the basis of our system.

Unlike the system developed by Siddharthan, our system attempts to correct sentences containing the pronouns *whom*, *whose*, and *that*, in addition to *who* and *which*. We are also interested in a general system that can handle a variety of structures in addition to the NP1 Prep NP2 structure and have modified the feature vector to account for them.

Another important difference between our automatic correction system and the work done by Siddharthan is that we will expand the scope of our study to look at the clause that follows the relative pronoun, enabling us to extract additional relevant features.

While the features used in Siddharthan's system are primarily based on semantic information obtained from Word Net regarding the

categories of noun phrases, syntactic knowledge will also be incorporated into our system. Using the UCREL CLAWS part-of-speech (POS) C5 tag set, a varying window of words before and after the relative pronoun will be tagged and used as attributes for the system. In doing so, we will determine if a system using semantic-based features in conjunction with syntactic features produces better results in correcting relative pronoun errors than one which relies exclusively on semantic features.

The final significant aspect of our learning system is that while the starting point of previous work has been grammatically correct English sentences, our system does not make that assumption. Our dataset consists of poorly translated English sentences, which, at minimum, contain one incorrectly translated relative pronoun. In many cases, the sentence contains other grammatical errors and we are attempting to correct the translation of the pronoun in the presence of these errors.

5. Experiments & Results

To train and test our system, a two-stage system was developed. The purpose of the first stage was to identify "bad" English sentences containing incorrect relative pronouns using a two-class framework where each instance was classified as good or bad. From here, the bad sentences were fed into the second stage, where the goal was to automatically correct the relative pronouns using syntactic and semantic features. For the first stage, the preliminary dataset of 72 instances described in Section 2 was used, along with the correct or "good" version of each sentence, for a total of 144 instances. The WEKA machine learning package, along with the Naïve-Bayes, Decision Tree, and One Rule algorithms were used to train and test the system.

An explanation of the six feature sets and the percentage of true negatives (out of a possible 72) for each can be seen in Table 2. Initially, 2/3 of the dataset was used for training and 1/3 for testing, along with cross-validation, where ten repetitions of the process were repeated. Although the results obtained with this evaluation method are more accurate, using 3-fold cross-validation

meant that the data was randomly split into 3 parts, resulting in too few instances to carry over to the second stage. Consequently, the entire dataset was used for testing and it is the true negatives that were identified in this experiment that were carried into the second stage. For each feature set, the lowest number of true negatives obtained from the three learning algorithms was used such that only the true negatives that all three algorithms had successfully identified were propagated to stage 2.

| Stage 1: Full dataset used for Testing | | | | |
|---|---|--------------|--------------|-----------|
| <i>Feature Set</i> | <i>Description</i> | <i>IR</i> | <i>DT</i> | <i>NB</i> |
| Stage1_1a | Word Net features only (See Annex A) | 84.72 | 72.22 | 84.72 |
| Stage1_2a | Word Net features, POS tag for (n+1) word | 84.72 | 88.89 | 84.72 |
| Stage1_3a | Word Net features, tags for n+1...n+5 words | 84.72 | 72.22 | 86.11 |
| Stage1_4a | Tags for n+1...n+5 words | 83.33 | 83.33 | 83.33 |
| Stage1_5a | Tags for n±1...n±5 words | 84.72 | 84.72 | 88.89 |
| Stage1_6a | Word Net features, tags for n±1...n±5 | 84.72 | 81.94 | 88.89 |

Table 2 - Stage 1 Experimental Results: Percentage of True Negatives

For the second stage, the system was expanded to a six-class problem, where the instances correctly identified as bad English sentences in stage 1 were then classified according to the correct relative pronoun. The results for stage 2 are given in Table 3.

| Stage 2 | | | |
|----------------------|-----------|-----------|-----------|
| Experiment #1 | | | |
| <i>Feature Set</i> | <i>IR</i> | <i>DT</i> | <i>NB</i> |
| Stage2_1a | 88.47 | 88.47 | 83.03 |
| Stage2_2a | 90.24 | 86.50 | 87.10 |
| Stage2_3a | 91.30 | 91.30 | 87.37 |
| Stage2_4a | 90.00 | 90.00 | 84.00 |
| Stage2_5a | 90.24 | 90.24 | 86.50 |
| Stage2_6a | 84.93 | 80.77 | 80.60 |
| Experiment #2 | | | |
| <i>Feature Set</i> | <i>IR</i> | <i>DT</i> | <i>NB</i> |
| Stage2_1b | 69.60 | 77.33 | 79.07 |
| Stage2_2b | 57.19 | 64.12 | 72.14 |
| Stage2_3b | 44.17 | 77.33 | 79.07 |
| Stage2_4b | 58.67 | 58.33 | 61.67 |
| Stage2_5b | 46.19 | 59.05 | 72.98 |

| | | | |
|-----------|-------|-------|-------|
| Stage2_6b | 34.17 | 63.87 | 76.50 |
|-----------|-------|-------|-------|

Table 3 - Stage 2 Experimental Results: Percentage of Correctly Classified Instances

The experimental results of stages 1 and 2 attempted to show several things. First, that a system is able to identify incorrect English sentences with fairly high accuracy and learn the correct relative pronouns for these sentences using syntactic descriptors in conjunction with semantic attributes. Based on the results obtained for experiment #1 of the second stage, it can be seen that the introduction of the POS tag feature for the five words following the relative pronoun in the third feature set produced the best results for the One Rule and Decision Tree algorithms. When the fourth and fifth feature sets were used, the second highest results were achieved, which is interesting because in this case, only syntactic features were used by the system. This suggests that relying solely on features based on Word Net categories may not be the best approach and that in order to achieve better results, other types of features must be incorporated as well.

The second important aspect demonstrated by the experimental results is that the system is capable of learning from incorrect instances. By including the incorrect relative pronoun feature, the system achieved fairly high results. However, this feature was removed in experiment #2 in order to eliminate the bias it introduces into the system. This bias relates to the specific machine translation system that was used. Because the language model of the system is built using specific sentence structures and templates, it is possible that certain relative pronouns will always be selected for specific phrase structures and thus, their occurrence may not be based on linguistic rules, but rather on pre-determined patterns. Although the results for the three algorithms on the six feature sets were considerably lower in this experiment, the percentage of correctly classified instances remained consistent for two of the three algorithms with the inclusion of syntactic features in the third feature set.

The overall results for the two-stage relative pronoun automatic correction system are summarized in Table 4.

| | |
|---|-------|
| % Correct sentences inappropriately identified as incorrect (stage 1) | 16.28 |
| % Incorrect sentences appropriately identified as incorrect (stage 1) | 83.72 |
| % Incorrect sentences corrected appropriately (stage 2) | 73.07 |
| % Incorrect sentences corrected inappropriately (stage 2) | 10.65 |

Table 4 – Summary of Results

6.0 Discussion & Future Work

At this point, there are several steps that must be taken in order to improve this automatic correction system. To begin, the dataset of French to English sentences must be increased to at least 250 examples since the accuracy of the first stage test results is proportional to the number of examples.

As described in Section 4, one of the most important aspects of our automatic learner is that it does not use grammatically correct sentences as the input to the system. An imminent task will be to determine whether the presence of these errors in the translated sentences contributes to the mistranslation of the relative pronoun. To do this, the translated sentences will be corrected so that every element except the relative pronoun is correct and represented using the feature vector described in Section 4. The experiments will be repeated to verify the validity of this hypothesis.

In terms of the overall functionality of the system, we are also interested in making it multilingual so translations from other languages such as Spanish to English will be considered as well. Once the system has been trained and tested, it will be evaluated using the BLEU evaluation technique.

7.0 Conclusions

The purpose of this research was to show that the correction of French to English relative pronoun translations is possible using Machine Learning methods. To accomplish this task, a preliminary data set of mistranslated sentences was

collected, each containing at least one incorrect relative pronoun. The representation of these sentences in feature vectors is rooted in the work of Siddharthan (2002) and expanded in several ways. First, the input to our learning system may or may not be a correct sentence as we have attempted to correct the relative pronouns in the presence of other errors. Second, by considering various clausal structures and relative pronouns, we have broadened the scope of study to include a wider range of corpora. Third, we have attempted to improve the baseline results by introducing syntactic features into the feature set. It is believed that by continuing to incorporate more semantic and syntactic features into the feature set, the baseline results may improve considerably.

REFERENCES

- G. Foster and R. Kuhn, "State-of-the-Art Statistical Machine Translation", in *TAMALE Seminar*, University of Ottawa, 3 February 2005.
- Brown et al., "A Statistical Approach to Machine Translation", *Computational Linguistics*, Volume 16, Number 2, June 1990, p.79-85.
- D. Marcu and W. Wong, "A Phrase-Based Joint Probability Model for Statistical Machine Translation", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, p. 133-139.
- K. Yamada and Kevin Knight, "A Syntax-based Statistical Translation Model", in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2001.
- http://www.gazette.uottawa.ca/index_f.php?newlang=fr_ench
- <http://babelfish.altavista.com/>
- <http://www.online-translator.com>
- A. Siddharthan, "Resolving Relative Clause Attachment Ambiguities using Machine Learning Techniques and Word Net Hierarchies", in *Proceedings of the 5th National Colloquium for Computational Linguistics in the UK (CLUK 2002)*, p.45-49.
- <http://www.comp.lancs.ac.uk/ucrel/claws/trial.htm>