TJHSST Computer Systems Project Proposal Development of a German-English Translator

Felix Zhang, Phil Graves

November 2, 2007

Abstract

Machine language translation as it stands today relies primarily on rule-based methods, which use a direct dictionary translation and at best attempts to rearrange the words in a sentence to follow the translation language's grammar rules to allow for better parsing on the part of the user. This project seeks to implement a rule-based translation from German to English, for users who are only fluent in one of the languages. For more flexibility, the program may implement limited statistical techniques.

Keywords: computational linguistics, machine translation

1 Introduction - Elaboration on the problem statement, purpose, and project scope

1.1 Scope of Study

I will focus on a rule-based translation system, because of time and resource constraints. I will start with part of speech tagging and lemmatization, and then progress to coding in actual grammar rules so that sentences can be parsed correctly, so that my program can handle more complex sentences as I embed more rules. At best, the program should be able to translate virtually any grammatically correct sentence, and find some way to resolve ambiguities. If I have time at the end of the year, I hope to try to implement

some basic statistical methods, such as part-of-speech tagging or single-word translation.

1.2 Purpose

The goal of my project is to use rule-based methods input to provide a translation from German in to English, or vice versa, for users who only speak one of the languages. Though the translation may be simple, the program still aids a user in that it provides a grammatically correct translation, which facilitates understanding of even primitive translations. Basic translations of short passages are especially helpful for users reading less formal text, as sentence structures tend to be less complex.

1.3 Expected results

When my project is finished, anyone can enter any German or English sentence, and verify that my translation is correct. I will compare my quality of translation to the quality of translation of a program based on statistical translation, and also commercial rule-based systems such as SYSTRAN.

1.4 Type of research

This project is use-inspired basic research. I want to create a rudimentary translation system, but also gain a grasp of a field in which I had no prior experience.

2 Background and review of current literature and research

Rule-based translation is the oldest form of language processing. A bilingual dictionary is required for word-for-word lookup, and grammar rules for both the original and target language must be hardcoded in to structure the output sentence and create a grammatical translation. Most online translators currently are based off of SYSTRAN, a commercial rule-based translation system. Statistiacl machine translation is the most-studied branch of computational linguistics, but also the hardest to implement. Statistical methods require a parallel bilingual corpus, which the program reads to "learn" the

language, determining the probably that a word translates to someting in a certain context. Statistical methods are considerably more flexible than rule-based translation, because they are essentially language-independent. Google Translate, which has access to several terabytes of text data for training, currently is developing beta versions of Arabic and Chinese translators based on statistical methods. Most research is being done with much more funding and resources than my project, and is thus much more advanced than my scope.

3 Procedures and Methodology

Input consists of a sentence in either English or German, and the output should be its equivalent translation, along with linguistic information for each word. The input requires that the sentence be grammatically correct, or it cannot be parsed correctly. The main components of my program will be the dictionary, part-of-speech tagging, morphological analysis, lemmatizing, and creating a parse tree. Part of speech tagging will be based on sentence position and capitalization, but I may later implement statistical tagging based on a pre-tagged corpus. Morphological analysis would use definite articles, suffixes, and adjective endings to determine linguistic properties such as gender, case, tense and person. Lemmatizing breaks down a word to its root form, to reduce the amount of entries required in the dictionary. The dictionary itself contains all pronouns and articles, along with a few verbs and nouns for testing purposes. The parse tree creates a tree based on the grammar of the base language, and rearranges it to accommodate the grammar rules of the target language.

4 Expected Results

The program, when complete, should be able to translate any given sentence in English or German into the target language. If I implement statistical methods successfully, the results could be useful to other seniors who create translation programs, since it would be the first to ever implement any statistical translation methods.