TJHSST Computer Systems Lab Senior Research Project
Development of a German-English Translator
2007-2008

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

A perfect machine translation of one language to another has never been achieved, because not all language expressions used by humans are grammatically perfect. It is also infeasible experimentally to code in every single grammar rule of a language. However, even a basic program that translates the basic idea of a sentence is helpfulfor understanding a text in a given language.

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.

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. Statistical 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 probability that a word translates to something 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 Development

The main components to a rule based translator are a bilingual dictionary, a part of speech tagger, a morphological analyzer that can identify linguistic properties of words, a lemmatizer to break a word down to its root, and a parse tree.

3.1 Dictionary

The dictionary stores a German word, its part of speech, its English translation, and any other data relevant to its part of speech, for example, for nouns, it also lists its plural form and gender. A large dictionary would be impractical for testing purposes, so I only include pronoun forms, conjunctions, and articles, with only a few nouns and verbs. These entries are stored in a hashtable.

3.2 Part of speech tagging

My part of speech tagger is based on sentence position and capitalization. These rules are specific to the language being translated. If a word is in between an article and a noun, it is an adjective. Hopefully, this system can be replaced by a statistical tagger, which would examine frequencies of tags appearing in a large tagged corpus, and would be considerably more flexible, because it would not depend on one specific arrangement of words in the input.

3.3 Morphological Analysis

Morphological analysis would use definite articles, suffixes, and adjective endings to determine linguistic properties such as gender, case, tense and person. It generates possible pairs of gender and case for nouns, and tense and conjugation for verbs. Two separate sets of pairs are generated for articles and modifiers, and the final list of possibilities is derived from the intersection of these two sets. This information is used for lemmatization.

3.4 Lemmatizer

A lemmatizer takes information from the morphological analysis and breaks a word down into its root form. For nouns, this means that plural nouns should be reduced to singular form, and suffixes resulting from different grammatical cases should be removed. For verbs, any ending from conjugation or tense should be removed. This saves considerable space in the dictionary, as I do not have to code in every inflected form of every word.

3.5 Parse tree

The parse tree arranges the sentence based on dependency grammar. Verbs connect from the subject to the direct object, and articles and adjectives are nodes of nouns. In translation, this tree must be rearranged to accommodate the target languages grammar.

3.6 Testing

Testing is conducted through input of sentences with new features. To test my lemmatizing component, I would input various inflected forms of a word to check the uniformity of the program's output. Varying sentence structures can also serve as a functional test to check the validity of newly-coded grammar rules in the parse tree.

4 Results

(projected) My program is able to translate a simple German or English sentence into the other language, provided the word is known in the lexicon. Most ambiguities in words are correctly resolved by a statistical tagger. The project fulfills its purpose as a simple translator with basic grammar rules, but would need an implementation of statistical methods to attain more flexibility in sentence structure parsing.