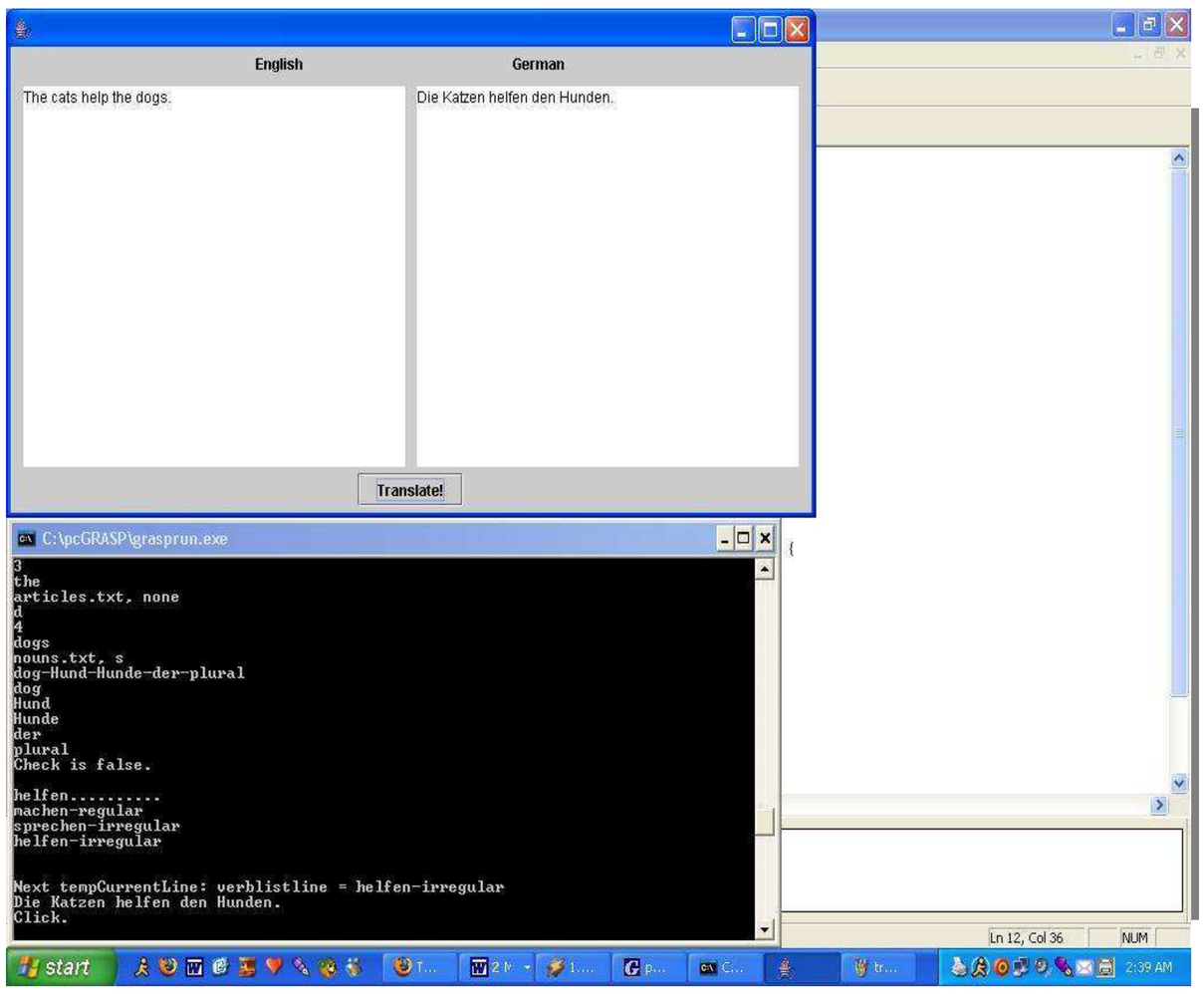
Natural Language Processing: Using Machine Translation in Creation of a German/English Translator

Jason Ji, 2004-2005

ABSTRACT - The field of machine translation - using computers to provide translations between human languages - has been around for decades. And the dream of an ideal machine providing a perfect translation between languages has been around still longer. This project attempts to take the beginning steps towards that goal, creating a translator program that operates within an extremely limited scope to translate between English and German. There are several different strategies to machine translation, and this project will look into them - but the strategy taken to this project will be the researcher's own, with the general guideline of "thinking as a human." For if humans can translate between language, there must be something to how we do it, and hopefully that something - that thought process, hopefully - can be transferred to the machine and provide quality translations.



BACKGROUND – The easiest translation method to code, yet probably least successful, is known as the direct approach. The direct approach takes the input language, performs morphological analysis - whereby words are broken down and analyzed for things such as prefixes and past tense endings, performs a bilingual dictionary look-up to determine the words' meanings in the target language, performs a local reordering to fit the grammar structure of the target language, and produces the target language output. This method is lacking in analysis of contextual information in a sentence.
The indirect approach involves some sort of intermediate representation of the source language before translating into the target language. In this way, linguistic analysis of the source language can be performed on the intermediate representation. The two main variants of the indirect approach are interlingua and transfer. The interlingua approach involves translating the source language into an intermediate language or representation that is not language dependent, and then translating into the target language without "looking back" at the source. The transfer method is similar, except that the transfer is language dependent - that is to say, the French-English intermediary transfer would be different from the English-German transfer. An interlingua intermediary can be used for multilingual systems.

Development - A sentence or multiple sentences are input into the translator, and a string consisting of that entire sentence (or sentences) is passed to the translate() function. The system loops through the string, finding all space (' ') characters and punctuation characters (comma, period, etc) and records their positions. The number of words in the sentence is determined to be the number of spaces plus one. By recording the position of each space, the string can then be broken up into the words. The start position of each word is the position of each space, plus one, and the end position is the position of the next space.

Two separate look-ups are performed in the process of translation upon the input of the source text and the breaking of it into each separate word. The first look-up finds the word in a general list, which contains the information to redirect the FileReader to a second list, which is part-of-speech specific. The word is found in the second list, and an entire line of semantic information related to the word is read-in along with the word. These information are stored together in an object of a subclass of the programmer-defined abstract Word superclass: if the word is a noun, then it and its information are stored in a Noun object, etc. The various Word-subclass objects are then stored in an array, on which corrections are performed for the case and gender of the nouns and conjugations of the verbs. The corrected Strings of the Word objects are then appended together and outputted to the text field.

Results -

Input	Output	Correct?
1) I see the dog.	Ich sehe den Hund.	Yes
2) You see the dog.	Du siehst den Hund.	Yes
3) The cats help the dogs.	Die Katzen helfen den Hunden.	Yes
4) The cats are the dogs.	<RUNTIME ERRORS>	No
5) Cats help dogs.	Katzen helfen Hunde.	No
6) The cats ate the pie.	Die Katzen essen die Torte.	Yes
7) I ate the pie.	Ich esse die Torte.	No
8) The cats speak with the dogs.	Die Katzen sprechen mit den Hunden.	Yes

1) Demonstrates verb conjugation and direct object case change.
2) Demonstrates verb conjugation beyond the first person.
3) Demonstrates verb conjugation in plural and dative plural case change.
4) Demonstrates an inability to work with irregular English verbs.
5) Demonstrates an inability to work without articles (though almost correct).
6) Demonstrates the usage of nominative and accusative plural correctly.
7) Demonstrates an inability to use the past tense, though conjugates correctly in the present tense.
8) Demonstrates usage of prepositions, changing the case correctly.

References

Hutchins, W.J. and Somers, H. L. An Introduction To Machine Translation. London: Academic Press, 1992.

German <-> English Dictionary. 2005. TU Chemnitz. November 2004 <http://dict.tu-chemnitz.de>.

LEO Deutsch-Englisches Wörterbuch. . LEO. November 2004 <http://dict.leo.org>.

“Machine Translation”. November 2004 <http://www.comp.leeds.ac.uk/ugadmit/cogsci/spchlan/machtran.htm>.