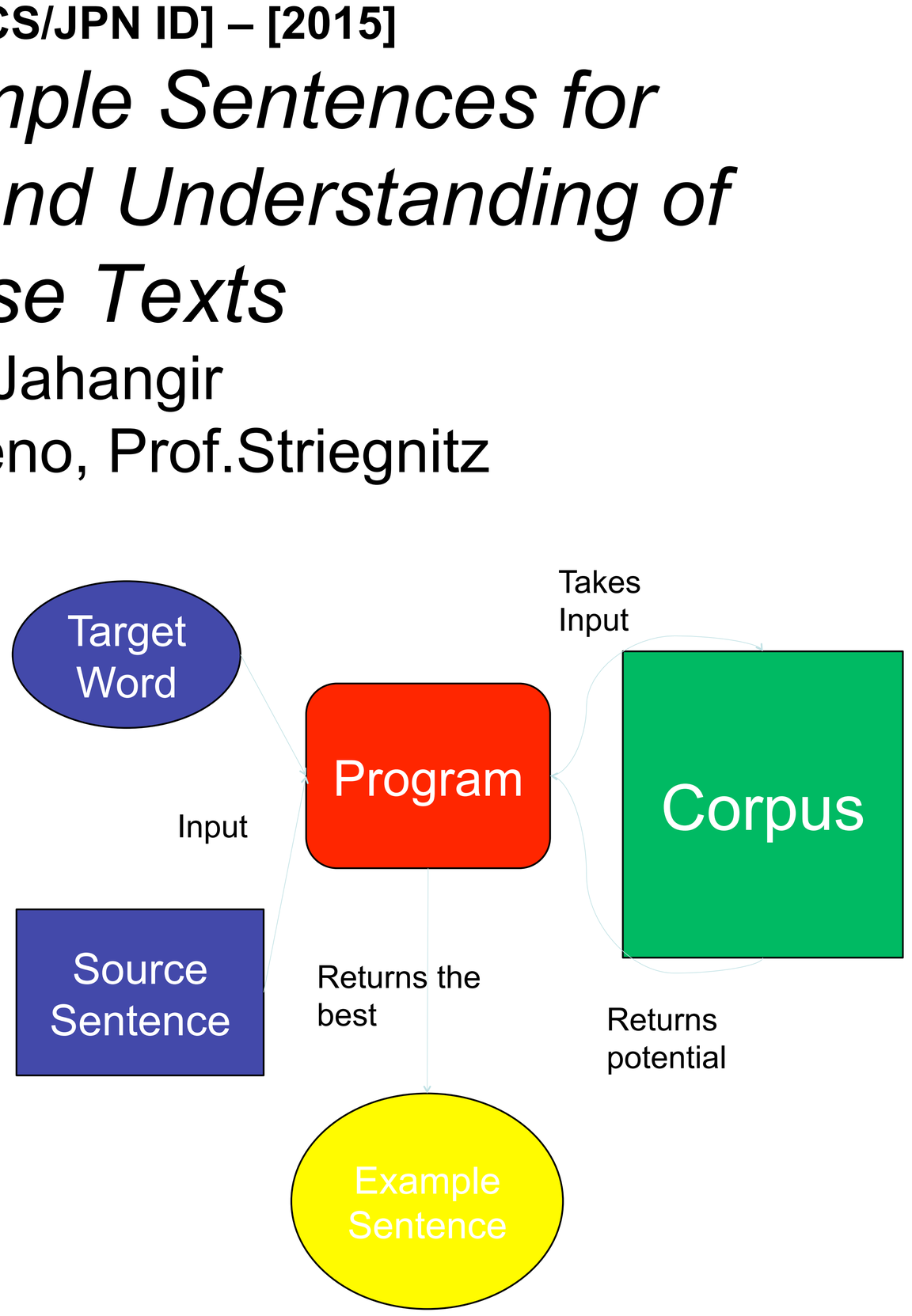
Senior Project – [CS/JPN ID] – [2015]

Extraction of Example Sentences for Improved Reading and Understanding of Japanese Texts

Nahian Jahangir
Advisor – Prof. Ueno, Prof.Striegnitz

Abstract

Learning languages can be a difficult task for everyone just starting. An especially difficult language to learn for native-English speakers is Japanese. Reading comprehension is a daunting task for novices who are unfamiliar with a language heavily dependent on context and uses three different writing systems. Therefore, I propose using example sentences for the reader that both improves their understanding of Japanese text. The qualifications of a good example sentence include retaining the context of the original sentence and having a suitable sentence appropriate for the level of the reader based on their knowledge of grammar and kanji.



Methods

I used the Tanaka Corpus, a multilingual parallel corpus, and the Simplified LESK algorithm, a word sense disambiguation algorithm, as my base. The corpus was removed of English and duplicate sentences, leaving 149,298 Japanese sentences. The LESK algorithm was also altered to calculate the word overlap of the source sentence and potential example sentences, rather than the word sense overlap. I evaluated the algorithm by looking at the scores the algorithm gave the sentences and how useful the sentence actually was. Improvements were made based on the analysis of the sentences and the score they received.

Project Goal

Use the Simplified LESK algorithm as a basis and continue to alter the algorithm to obtain helpful example sentences from the Tanaka Corpus.

Initial Conclusions

Method 1
"面倒","両親が年をとったら面倒を見るつもりです“

RANKING:
sentence: 私は面倒なことになると予想した。score: 0.78
sentence: 私たちは代わる代わる子供たちの面倒を見た。score: 0.78
sentence: 面倒なことになるよ。score: 0.75
sentence: 私が面倒をみます。score: 0.75
sentence: 面倒だな。score: 0.67
sentence: 面倒が起こるのではないかと私は恐れている。score: 0.67
sentence: 放課後に面倒なことが起きた。score: 0.67
sentence: 心配しないで、お前の面倒は見るから。score: 0.67
sentence: 彼女は父親が死ぬまで面倒を見た。score: 0.63
sentence: 彼女は家で子供の面倒を見ているよ。score: 0.63

This method was based on the collocation of the potential example sentences. This is the top ten ranking of example sentences the algorithm calculated. The last two highlighted red on the list were evaluated as good example sentences by myself and Professor Ueno.

Evaluations and Discussion

This method involved creating a dictionary of words of all the sentences that had the target word and acquiring the top 20 words that had the highest frequency within the corpus. Using the set of 20 words, I found the top 10 sentences that had the highest intersection with that set. This method also included the normalization of scores and the removal of grammatical particles within the sentences. Results showed that 2/10 sentences of the top 10 were helpful, which shows improvement from the baseline of scoring based on word overlap. Future work involves continuing further analysis of the sentences and alterations to the algorithm.