Beam Search Algorithm for Statistical Machine Translation

compiled by D. Gueorguiev, 8/16/2025

We are going to discuss a search procedure for Statistical Machine Translation (SMT).

We will use the trigram language model for simplicity. We would like to focus on the most likely translation hypotheses during the search process. For this purpose, we will use a data-driven beam search algorithm which is the subject of this document.

The search procedure is based on a Dynamic Programming Algorithm (DPA) to solve the Traveling Salesman Problem (TSP). The cities in the TSP correspond to source positions of the input sentence. By imposing judiciously defined constraints on the word ordering DPA becomes more effective as the number of possible reordering is significantly reduced. A scheme devising a set of general reordering constraints is proposed where each constraint is expressed in terms of combinatorial restrictions on the processed sets of source sentence positions. A set of four parameters is given to control the word reordering. Additionally, a set of four states is introduced to deal with grammatical reordering restrictions (e.g. for translation direction German-to-English, the word order difference is mainly due to the German verb group. In combination with the reordering restrictions a data-driven beam search organization for the search procedure is proposed.

A source string

is to be translated into a target string

Then

# References

[1] [Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation, Christoph Tillmann, Hermann Ney, 2003](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/Word_Reordering_and_a_Dynamic_Programming_Beam_Search_Algorithm_for_Statistical_Machine_Translation_Tillman_IBM_2003.pdf)

[2] [Language Translation Apparatus and Method Using Context-Based Translation Models, Adam L. Berger et al, US5,510,981, 1996](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/Language_translation_apparatus_and_method_of_using_context-based_translation_models_US_Patent_5510981.pdf)

[3] [The Mathematics of Statistical Machine Translation: Parameter Estimation, Peter F. Brown et al, IBM T.J.Watson Research Center, 1993](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/The_Mathematics_of_Statistical_Machine_Translation-Parameter_Estimation_Brown_1993.pdf)

[4] [A Statistical Approach to Machine Translation, Peter F. Brown et al, IBM T.J.Watson Research Center, 1990](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/A_Statistical_Approach_to_Machine_Translation_Brown_1990.pdf)

[5] [A DP-based Search Algorithm for Statistical Machine Translation, S. Niessen et al, Aachen U. of Technology, 1999](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/A_DP_based_Search_Algorithm_for_Statistical_Machine_Translation_Niessen_1999.pdf)

[6] [Statistical Machine Translation, Final Report, JHU Workshop, Yasser Al-Onaizan et al, 1999](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/Statistical_Machine_Translation_FInal_Report_JHU-1999-AlOnaizan.pdf)

# Appendix

## Introductory Notes to Statistical Translation

A string of English words , , can be translated into a string of French words in many different ways. Knowing the broader context in which occurs may allow us to reduce the set of acceptable French translations. In general, we will assume that every French string, , is a possible translation of with some non-zero probability. We assign to every pair of strings a number . Given a French string of words , we want to find the string which is the best English translation. We minimize the error which we could make by choosing that English string for which is largest.

Obviously:

(smt.1)

Noting that is independent on , we will estimate by selecting that which will maximize the product .

Thus,

(smt.2)

Question: Can we construct approximations to the distributions and that are good enough for acceptable quality of the translation?

Challenges:

1 ) **the language modeling problem**: estimate the *language model probability*

2 ) **the translation modeling problem**: estimate the *translation model probability*

3 ) **the search problem**: devise and implement effective suboptimal search for the English string that maximizes the product of the two probability models

Question: why do we estimate and instead of ?

To answer this question, we divide French and English strings into sets of those that are well-formed and those that are ill-formed. We denote those with , , , accordingly. For the translation model we want to assign meaningful probability values to as much as possible of . That is , we want to be defined on as large subset of as possible. Note that it is not as important that the model for covers as large subset of as possible.