Parameter estimation for Statistical Machine Translation

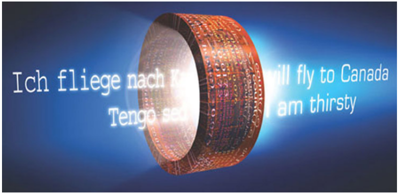
Presenter : Amrutha Varshini Ramesh

Department of Electrical & Computer Engineering
University of Illinois at Chicago

CS531, Detection and Estimation Theory
Final Project presentation

Outline

Problem at hand - Machine Translation



Problem at hand

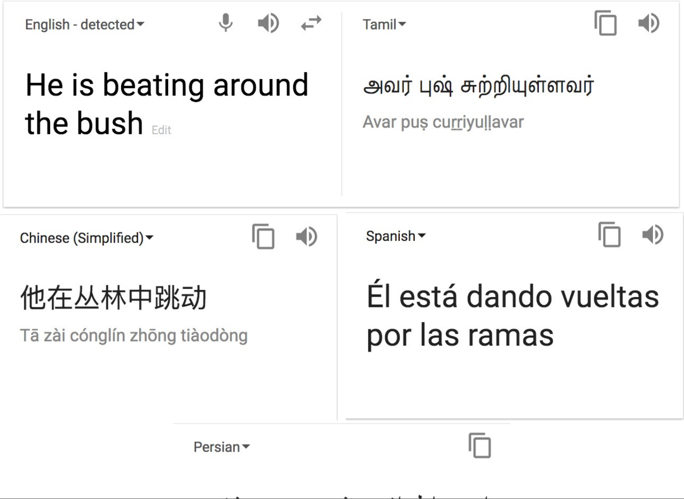
Translate sentence s_1 from language l_1 to s_2 from language l_2

- l_1 : Source language
- l_2 : Target language
- s_1 : Sentence in source language
- s_2 : Sentence in target language
- ws_i : A Word in s_1
- wt_i : A Word in s_2

Outline

Challenges in Machine Translation

- Changes in the language structure
- Identifying compound words and two-word verbs
- Words with multiple meanings, Idioms and metaphors !!! This one is tricky !!



Solving the Machine Translation problem

Two problems to solve

- Given s_1, find best translations for each word in the target language
- Given these translations, find the right alignment

In most practical Machine Translation systems, we need to solve both the problems!

Classes of approaches

Classes of approaches

- Rule-based approach
- Corpus-based approach

Rule-based approach

- Uses manually curated rules for translation.
- Based on the morphological, syntactic and semantic map between the two languages.
- Laborious and time-consuming.

Corpus-based approach- Statistical Machine Translation

- Data driven methods that learn the rules of translation automatically from the corpus(data).
- A Statistical model is constructed who parameters are estimated from the corpus.
- Translations for unseen sentences are based on the learned model(s)

Outline

Statistical Machine Translation(SMT)

SMT model

$$\underset{s_2 \in l_2}{\operatorname{argmax}}\, p(s_2|s_1) = \underset{s_2 \in l_2}{\operatorname{argmax}} \underbrace{p(s_1|s_2)}_{\text{Translation model}} * \underbrace{p(s_2)}_{\text{Language model}}.$$

Translation model:

- $p(s_1|s_2)$ is the conditional prob. of the source sentence given the target sentence
- Likelihood of the source sentence s_1 given target s_2.
- Why are we modeling this? Think discriminative power of the model.

Language Model:

- $p(s_2)$ is the marginal prob. of the target sentence generated.(Think correctness of the sentence)
- Gives us how probable a given target sentence s_2 is, under the rules of the target language.

SMT - Translation model

- Models the relationship between source and target sentences.
- Harder part of the translation mechanism because of the number of unknowns as it models both translation and alignment.

How to estimate $p(s_1|s_2)$

- As discussed earlier, it's a 2 stage process.
- Stage 1: Given $ws_i \in s_1$ find the best $wt_j \in s_2$. This is word-by-word translation.
- Stage 2: Given the word-by-word translations, find the best alignment(or mapping) between $ws_i's$ and $wt_j's$.

Chicken and egg problem! We don't know the solution to either of the stages to begin with. EM!

Outline

Expectation Maximization(EM) for SMT

Stage1 - Expectation

Compute $p(a|s_1, s_2)$

Learning the best alignment function a that maps words ws_i to wt_j

Stage2 - Maximization

$wt_j = \text{argmax}_{wt \in l_2}\, p(w_t|ws_i)$

Given a training corpus and an alignment(mapping of word positions), estimating this is trivial(Maximization). Here it's just **MLE**.

- This falls into the framework of EM.
- The detailed algorithm is discussed later.
- Basic idea is to alternate between estimating optimal translations between words in the two sentences and finding the best possible alignment.

EM algorithm for SMT

Algorithm 1 EM Algorithm for SMT

Input: Set of sentence pairs (s_2, s_1)
Output: Translation prob. $p(wt_i|ws_j)$

initialize $p(wt|ws)$ uniformly
// *initialize*
while not converged **do**
  count$(wt_i|ws_j) = 0$ **for all** i, j
  total$(ws_j) = 0$ **for all** j
  for all sentence pairs (s_1, s_2) **do**
    // *Compute normalization*
    for all words wt_i in s_2 **do**
      s-total$(wt_i) = 0$
      for all words ws_i in s_1 **do**
        s-total$(wt_i) += p(wt_i|ws_j)$
      end for
    end for
  end for
  // *estimate probabilities*
  for all target words wt_i **do**
    for all source words ws_j **do**
      $p(wt_i|ws_j) = \frac{\text{count}(wt_i|ws_j)}{\text{total}(ws_i)}$
    end for
  end for
end while

Outline

Results

Dataset descriptions

English-French Europarl dataset

# sentences	50000
# words in the vocabulary	37000
Avg. sent. length	6

English-Spanish Europarl dataset

# sentences	50000
# words in the vocabulary	74000
Avg. sent. length	5

Performance statistics

Table: Performance of EM-SMT on English to Spanish Europarl dataset

Precision	0.596
Recall	0.487
F1 score	0.536

Translated sentences

- s_1: De région en région, les situations sont très, très différentes
  s_2: The situation varies to an enormous degree throughout the regions.
  Google translate: From region to region, situations are very, very different
- s_1: Je ne le crois pas.
  s_2: I do not believe so
  Google translate: I do not believe that
- s_1: La procédure a connu quelques ralentissements au niveau du Conseil, ralentissements dus notamment à des divergences de vues concernant l' accord sur la libre circulation des personnes
  s_2: The procedure has undergone some delays in the Council due, in particular, to differences of opinion regarding the free movement of persons.
  Google translate: The procedure has seen some slowdowns at Council level, slowdowns due in particular to differences of opinion concerning the agreement on the free movement of persons

Possible improvements to the existing model

- Recall that we had two estimation problems in the previous EM procedure.
- A Translation model : Translations of words in sentence s_1 to words from language l_2.
- An Alignment model : Models the alignment between the translated words.
- The current translation model is *Naive.* Simple MLE.
- We could incorporate richer translation models, that incorporates additional properties of the words in both the languages. For instance,
  - We could incorporate history(Markov models).
  - We could use grammatical structures in both the languages and model them explicitly(for example : POS tags, coreference structures).
  - Additional word related features - capitalization, word positions etc

Richer Translation Model

- Basic idea here is build a model that incorporates arbitrary features of words in both the sentences.

Consider the sentence : "He went to school yesterday". Consider the first word "He". For a typical language model, the following additional features could be useful for predicting the right translation.
Is_Capitalized : Yes, : Is_first_word: Yes, Is_verb : No

- This makes the model richer and easens the burden on the alignment problem.
- How to incorporate arbitrary features in our Translation model?
- There are sophisticated class of Machine Learning models called *Conditional Random Fields(CRF)*, that let's us incorporate additional(arbitrary) features to words apart from just the word itself.

Translation model with Deep Networks

- Deep Neural network(DNN) models such as RNN and LSTM can model sequential structure
- Model setup is simple and DNN's are highly sophisticated models.
- I'm currently experimenting with RNN's.
- Results will be added in the report.

References I

- Statistical-Machine-Translation/tp3.sujet.pdf at master · Mandarancio/Statistical-Machine-Translation · GitHub.
- Chris Callison-Burch.
  Machine translation: Word-based models and the EM algorithm Chris Callison-Burch Word-based translation models and EM.
  2007.
- Andrej Karpathy.
  The Unreasonable Effectiveness of Recurrent Neural Networks, 2015.
- Philipp Koehn.
  Europarl: A Parallel Corpus for Statistical Machine Translation.
- Thomas Lavergne, Josep Maria Crego, Alexandre Allauzen, and François Yvon.
  From n-gram-based to CRF-based Translation Models.
  pages 542–553.

References II

Adam Lopez.
A Survey of Statistical Machine Translation.
2007.

Charles Sutton and Andrew Mccallum.
An Introduction to Conditional Random Fields.
Machine Learning, 4(4):267–373, 2011.

Chong Wu, Can Yang, Hongyu Zhao, and Ji Zhu.
On the Convergence of the EM Algorithm: A Data-Adaptive Analysis.
nov 2016.

Questions?

Thank You!