Word Based Model: IBM Model 1

Ayushi Dalmia Natural Language Application CSE573 201307565

Introduction

The Internet contains billions of web pages, as they come in all kinds of languages, a great deal of information is not available to us. A practical application would be a browser that translates these pages in a preferred language. In the field of statistical machine translation (SMT), we try to build algorithms to translate from one language to the other by mere statistics taken from large bi-text corpora. When we adopt the SMT approach, we represent all individual or groups of words (cepts) as having a connection to zero, one or many foreign cepts under a probability value. This means that both the alignments and the probabilities need to be extracted from the bi-text. The main problem here is that we need the alignments to estimate the probabilities and the probabilities to estimate the alignments. These kinds of problems can be solved with the Expectation Maximisation (EM) algorithm.

In this project, the word based model implemented. A word based model is a simple model for machine translation that is based solely on lexical translation, the translation of words in isolation. This requires a dictionary that maps words from one language to another. Using the bi-text corpus we train the model and calculate the translational probability deploying the principle of EM.

Algorithm

In this model the EM Algorithm is used to calculate the translational probabilities between words. The translation probabilities are initialised uniformly. In the E step, the count of each word is done for its possible translated word. In the M step we update the translation probabilities using the counts from the E step. . Under IBM model I, we assume that each target word is to be generated by exactly one source word, which can also be the null word. The pseudo code for the algorithm is given by Figure 1.

```
Input: set of sentence pairs (e, f)
Output: translation prob. t(e|f)
initialize t(e|f) uniformly
while not converged do
  // initialize
  count(e|f) = 0 for all e,f
  total(f) = 0 for all f
  for all sentence pairs (e,f) do
    // compute normalization
    for all words e in e do
      s-total(e) = 0
      for all words f in f do
        s-total(e) += t(e|f)
    end for
  end for
  // collect counts
  for all words e in e do
    for all words f in f do
      count(e|f) += t(e|f) / s-total(e)
      total(f) += t(e|f) / s-total(e)
    end for
  end for
  // estimate probabilities
  for all foreign words f do
    for all English words e do
      t(e|f) = count(e|f) / total(f)
    end for
  end for
end while
```

EM training algorithm for IBM Model 1.

Figure 1

The convergence of the problem is determined on the basis of the perplexity. Given *s* sentences, the perplexity for a model with translational probabilities *t (e|f)* is given by,

$$\log_2 PP = -\sum_s \log_2 p(e_s|f_s)$$

Here *e* is the sentence in the target language and *f* is the sentence in the source language. The probability for each sentence is calculated using the following:

$$p(e|f) = \frac{\epsilon}{l_f^{\,l_e}} \sum_{j=1}^{l_e} \sum_{i=1}^{l_f} t(e_j|f_i)$$

where l_e is the number of words in the target sentence, l_f is the number of words in the source sentence and t (e_j | f_j) is the translation probability of the target word e_i given the source word f_j .

Ideally, we converge when the perplexity of the current iteration is same as the perplexity of the previous iteration

Dataset

In order to implement IBM Model 1, we first need to train the model to find out the translational probabilities. In order to do so, we used a parallel corpus of English and German provided in the paper Koehn, Philipp. "Europarl: A parallel corpus for statistical machine translation "*MT summit*. Vol. 5. 2005. The dataset consists of 1920209 lines in both the corpus. From this corpus we experiment using 1,000, 10,000 and 20,000 sentences for both English and German.

Experimental Results

Implementation of IBM Model 1 was done as described in Figure 1. Random selection of 1,000, 10,000 and 20,000 sentences is done for training the model. For each set of training sentences the EM algorithms is applied and the translation probabilities are calculated. Convergence is determined on the basis of perplexity values. In the project we use the translation probabilities obtained after 10 iterations only. The experiment is done in both the direction, i.e. from English to German and from German to English.

The alignment of sentences is found using the translation probabilities obtained after training the model with 20,000 sentences.

Perplexity

The perplexity is calculated for 10 iterations using different number of sentences. It is calculated for translation in both directions. The values are rounded off up to 4 decimal places. Table 1 represents the perplexity values for German to English translation for the 1st 10 iterations. Similarly, the perplexity values for English to German translations was calculated as shown in Table 2.

Iteration	Number of Sentences for Training		
	1000	10000	20000
1	5.3932E+58437	5.6827E+710385	1.6671E+1501699
2	1.3772E+58381	3.7445E+707903	2.3365E+1499451
3	1.8422E+58323	6.7413E+706311	2.5704E+1497121
4	2.6734E+58285	4.7232E+705910	2.5704E+1495926
5	1.8142E+58266	5.2182E+705823	6.6001E+1495510
6	4.7377E+58257	7.1791E+705733	1.9546E+1495414
7	1.4092E+58254	9.1175E+705623	3.0018E+1495395
8	7.9355E+58252	3.9410E+705552	8.4380E+1495325
9	4.6027E+58252	1.2809E+705432	9.3986E+1495262
10	3.5017E+58252	4.0721E+7055326	2.1561E+1495162

Table 1: Perplexity Values for training from German to English

Iteration	Number of Sentences for Training		
	1000	10000	20000
1	1.4197E+68481	2.2492E+830619	3.7094E+1735712
2	3.7015E+68404	1.0399E+830822	1.0809E+1689119
3	4.1833E+68306	1.8487E+827344	9.6479E+1676556
4	2.7228E+68240	1.2891E+827340	1.8382E+1673537
5	8.4447E+68210	3.0998E+827201	1.7027E+1670527
6	6.1438E+68199	3.3423E+826996	9.0185E+1667437
7	6.5502E+68195	4.8840E+826925	1.2577E+1665352
8	3.0147E+68194	8.6640E+826856	4.6145E+1663258
9	1.3271E+68194	1.0714E+826833	2.4312E+1661456
10	1.4021E+68194	6.9537E+829061	1.7086E+1659836

Table 2: Perplexity Values for training from English to German

Alignments

The alignment for the sentences is found using the translation probabilities. The translation probabilities are calculated using 20,000 sentences from training and after 10 iterations. The alignment for both the training and testing set is deduced.

Samples from Training Set

The alignments for English to German translations are as follows:

English: I will not talk about codecision here, but this certainly does not mean I do not consider it important.
German: Ich möchte hier nicht auf die Frage der Mitentscheidung eingehen, jedoch nicht, weil ich sie nicht für wichtig hielte.
Translated: mindestsozialhilfesatzes agrarhöchstbetrag entwicklungsprojekte all wochen zweifellos abwertung entwicklungsprojekte entwicklungsprojekte heute mitgliedstaates entwicklungsprojekte geschäft mindestsozialhilfesatzes agrarhöchstbetrag entwicklungsprojekte energie entwicklungsprojekte entwicklungsprojekte

English: Judging from the discussions held over the past five or six months, this could well be the case.
German: Wenn ich mir die Diskussionen der letzten fünf, sechs Monate vergegenwärtige, wäre das durchaus möglich gewesen.

Translated: möglich pluralismus kommen wochen ganz wochen kommen wochen wochen inhaftierten wochen wochen entwicklungsprojekte wochen wochen mindestsozialhilfesatzes kommen mit

English: Yet, every year we have the same problem**.**
German: Das Problem ist doch jedes Jahr das gleiche.
Translated: wochen pluralismus wochen pluralismus entwicklungsprojekte kommen wochen mitgliedstaates

English: The Leader III scheme is set to be implemented some time later this year.
German: Die Durchführung von Leader III ist für das laufende Jahr vorgesehen.
Translated: kommen ganz laufende einhaltung es wochen kommen mindestsozialhilfesatzes all wochen wochen fischerei entwicklungsprojekte wochen

English: What guarantees do we have that they will grow into first-class players in the league of the internal market?
German: Welche Garantien haben wir, daß sie sich zu Topspielern in der Liga des Binnenmarkts entwickeln?
Translated: wochen freiheiten agrarhöchstbetrag pluralismus entwicklungsprojekte mindestsozialhilfesatzes entwicklungsprojekte agrarhöchstbetrag irische pluralismus NULL sollte eine kommen antrags der kommen schengen arbeitsrechtlichen

The alignments for German to English translations are as follows:
German: Vielen Dank, Herr Poettering.
English: Thank you, Mr Poettering.
Translated: sector all secondly just

German: Wir müssen zugeben, daß die EU die Entwicklung der ärmeren Länder bereits - nach meinem Dafürhalten sogar sehr großzügig - unterstützt hat.
English: I remember what Portugal and Greece used to be like when I drove through those countries for the first time twenty-five years ago.
Translated: we aided das secondly stock limited stock secondly raining gdp limited secondly limited represent sector secondly limited all secondly four

German: Ich akzeptiere nicht, daß ein Umsetzungsproblem in der Praxis der Grund für eine Änderung der Rechtsordnung sein soll.
English: I do not accept that a practical transposition problem should give rise to changing the law.
Translated: four all cannot secondly elvs a foul raining secondly raining secondly four foul secondly raining office secondly secondly

German: Wir brauchen in dieser Frage wesentlich mehr Offenheit und Transparenz als bisher.
English: There needs to be much more openness and transparency on this issue than there has been to date.
Translated: we secondly foul paperless secondly all paperless atmosphere stock represent four secondly

German: Der Bericht befaßt sich mit der Harmonisierung von Prüfungsanforderungen für Sicherheitsberater, die im Bereich der Beförderung gefährlicher Güter auf Straße, Schiene oder Binnenwasserstraße tätig sind.

English: The report looks at the issue of harmonising the examination requirements for safety advisors working in the areas of transportation of dangerous goods by road, rail and inland waterway.
Translated: raining limited all four four raining all foul and four and stock four stock raining sector restriction union's four all and limited and breadth four

Samples from Testing Set

The alignments for English to German translations are as follows:

English: Resumption of the session
German: Wiederaufnahme der Sitzungsperiode
Translated: wie der kommen morgige

English: Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.
German: Wie Sie feststellen konnten, ist der gefürchtete "Millenium-Bug " nicht eingetreten. Doch sind Bürger einiger unserer Mitgliedstaaten Opfer von schrecklichen Naturkatastrophen geworden.
Translated: eingangs entwicklungsprojekte wochen agrarhöchstbetrag entwicklungsprojekte wochen kommen NULL morgige europäer unseres kommen zustande wochen kommen wochen eine eine wochen der entwicklungsprojekte all eine berichterstatter der einhaltung übergegangen mindestsozialhilfesatzes rom wochen dafür

English: You have requested a debate on this subject in the course of the next few days, during this part-session.
German: Im Parlament besteht der Wunsch nach einer Aussprache im Verlauf dieser Sitzungsperiode in den nächsten Tagen.
Translated: wochen entwicklungsprojekte all eine wochen entwicklungsprojekte entwicklungsprojekte wochen eine kommen ministerrates der kommen wochen wochen wochen mitgliedstaates entwicklungsprojekte unterausschuß

English: In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union.
German: Heute möchte ich Sie bitten - das ist auch der Wunsch einiger Kolleginnen und Kollegen -, allen Opfern der Stürme, insbesondere in den verschiedenen Ländern der Europäischen Union, in einer Schweigeminute zu gedenken.
Translated: eine kommen fälle mindestsozialhilfesatzes dienste wochen kommen unseres eine sprechen arbeitsrechtlichen sprechen entwicklungsprojekte eine wochen der unterausschuß entwicklungsprojekte all entwicklungsprojekte wochen der abgehalten kommen afghanistans wochen unterwerfung wochen der kommen ernähren kann eine kommen wochen entwicklungsprojekte der kommen entwicklungsprojekte dienste

English: Please rise, then, for this minute' s silence.
German: Ich bitte Sie, sich zu einer Schweigeminute zu erheben.
Translated: rilke wochen agrarhöchstbetrag entwicklungsprojekte entwicklungsprojekte sprechen arbeitsrechtlichen sprechen

The alignments for German to English translations are as follows:

German: Wiederaufnahme der Sitzungsperiode
English: Resumption of the session
Translated: sector increase September

German: Wie Sie feststellen konnten, ist der gefürchtete "Millenium-Bug " nicht eingetreten. Doch sind Bürger einiger unserer Mitgliedstaaten Opfer von schrecklichen Naturkatastrophen geworden.
English: Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.
Translated: unimaginative secondly represent all is increase NULL NULL reallocations and limited unimaginative secondly all secondly secondly trivialisation four restore summer all

German: Im Parlament besteht der Wunsch nach einer Aussprache im Verlauf dieser Sitzungsperiode in den nächsten Tagen.
English: You have requested a debate on this subject in the course of the next few days, during this part-session.
Translated: four secondly secondly increase secondly secondly secondly all four operations limited september four four represent impression

German: Heute möchte ich Sie bitten - das ist auch der Wunsch einiger Kolleginnen und Kollegen -, allen Opfern der Stürme, insbesondere in den verschiedenen Ländern der Europäischen Union, in einer Schweigeminute zu gedenken.
English: In the meantime, I should like to observe a minute's silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union.
Translated: schengen secondly systematic secondly secondly the is unimaginative increase secondly all secondly four secondly secondly impression increase NULL secondly four four all represent increase secondly secondly four secondly and raining trivialisation.

German: Frau Präsidentin, zur Geschäftsordnung.
English: Madam President, on a point of order.
Translated: raining all secondly all

Analysis

It is found that with every iteration the value of perplexity decreases and moves towards convergence. Due to a very small training set, the probability values are very small and hence taking the logarithm results in very large values for the perplexity. It is interesting to note that the value of perplexity decreases with every iteration. This indicates that eventually the algorithm will converge and the translation probability values will become constant.

From the alignment it is found that since this model takes the greedy approach by considering the word with the highest probability, the translations are found not to up to the mark. Also, it does not takes into consideration the possibility of phrase translation and the context at all.

Although an attempt was made in the experiment to estimate the error, but after obtaining the alignments the idea was dropped. The translations obtained is based merely on a greedy approach completely ignoring the grammatical structure of the sentence. Also, the number of words in the original sentence is different from the number of words in the final sentence. Hence, error analysis on

the translated sentence seemed infeasible. There was no way that we could check and evaluate how good the translations are!

IBM model 1 has some widely known structural limitations such as only supporting a many to one and not a many to many alignment. It completely ignores the positions of the words in the sentences and any other alignments already set. The model is very weak in terms of reordering, as well as adding and dropping words. As stated in Koehn "*According to IBM Model 1, the best translation for any input is the empty string"*