Natural Language Application – CSE573 : Assignment 1

Objective:

1. Train the IBM Model 1 with different sizes of the corpus: 1000, 10000 and 20000 sentence pairs.
2. Report the perplexity of your model on both directions for 10 iterations. (6 tables in total)
3. Report the most likely alignments for 10 sentences (5 from your training set, and 5 from outside our training set), in both directions.
4. Discuss the errors generated by this model

Approach:

The following pseudocode has been used in order to implement IBM Model 1 in **Python**.

```
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
```

- The model has been trained with the corpuses of size 1000, 10000 and 20000. Following files are used:
  - de1000.txt, en1000.txt
  - de10000.txt, en10000.txt
  - de20000.txt, en20000.txt

NOTE: The given perplexity values are given as natural logarithm of actual perplexity. Due to language number limitations values could not be obtained. These values can be considered actual values since they are directly proportional to the actual 2^x values.

Following tables display the perplexity obtained during each iteration of EM training for corpus of size 1000:

	Source: German, Target: English (log PP)
Iteration 1	169116.359935
Iteration 2	159599.681391
Iteration 3	154840.903686
Iteration 4	152109.921648
Iteration 5	150452.343208

Iteration 6	149408.382893
Iteration 7	148727.14866
Iteration 8	148268.032648
Iteration 9	147949.601494
Iteration 10	147722.944155

	Source: English, Target: German (log PP)
Iteration 1	155844.45246
Iteration 2	145857.595364
Iteration 3	140913.978985
Iteration 4	138086.51278
Iteration 5	136350.630922
Iteration 6	135245.433336
Iteration 7	134518.427934
Iteration 8	134025.45148
Iteration 9	133681.886041
Iteration 10	133436.43619

Following tables display the perplexity obtained during each iteration of EM training for corpus of size 10000:

	Source: German, Target: English (log PP)
Iteration 1	1856428.08748
Iteration 2	1704424.70741
Iteration 3	1635842.23066
Iteration 4	1603853.75469
Iteration 5	1587677.12205
Iteration 6	1578614.00038
Iteration 7	1573093.7161
Iteration 8	1569513.53318
Iteration 9	1567077.25937
Iteration 10	1565355.28667

	Source: English, Target: German (log PP)
Iteration 1	1685379.23535
Iteration 2	1533508.76392
Iteration 3	1463796.96165
Iteration 4	1429982.9579

Iteration 5	1412665.71519
Iteration 6	1403024.20756
Iteration 7	1397201.20179
Iteration 8	1393442.01098
Iteration 9	1390887.07461
Iteration 10	1389079.43655

Following tables display the perplexity obtained during each iteration of EM training for corpus of size 20000:

	Source: German, Target: English (log PP)
Iteration 1	3856398.93869
Iteration 2	3509645.04754
Iteration 3	3360487.62869
Iteration 4	3296298.07075
Iteration 5	3265350.80734
Iteration 6	3248393.39762
Iteration 7	3238163.98849
Iteration 8	3231559.32596
Iteration 9	3227075.71898
Iteration 10	3223909.14913

	Source: English, Target: German (log PP)
Iteration 1	3487131.11031
Iteration 2	3144958.14447
Iteration 3	2993396.39406
Iteration 4	2925180.26033
Iteration 5	2892095.36384
Iteration 6	2874228.08664
Iteration 7	2863600.94217
Iteration 8	2856783.09087
Iteration 9	2852154.36299
Iteration 10	2848874.41532

Alignment of 5 sentences using obtained model when source is German and target is English (Training sentences used)

1. Frau Präsidentin, zunächst möchte ich Herrn Koch für seinen Bericht danken, der die Verkehrssicherheit zum Thema hat.

   Best Alignment
   should Madam Madam heart, like I Mr Koch for his report thank of which heart, heart, issue heart,

2. Sie haben sie auch so beantwortet, wie ich wußte, daß Sie sie beantworten würden.

   Best Alignment
   would you knew kind, Indeed kind, responded as I knew knew you kind, responded kind,

3. Hier besteht jetzt dringender Handlungsbedarf.

   Best Alignment
   need urgency urgency now. urgency urgency

4. Herr Präsident, auch ich stimme zu, daß der hier von uns behandelte Bericht nur ein kleiner Schritt - ein noch zu kleiner Schritt - zur Schaffung jenes gemeinsamen Raums der Freiheit, der Sicherheit und des Rechts darstellt, den Europa hoffentlich eines Tages verkörpern wird.S

   Best Alignment
   is Mr President, too I agree agree that of become. of we become. become. examining one become. step - one still too become. step - towards become. become. common freedom, of freedom, of security and construction freedom, become. the Europe hopefully construction day become. become.

Alignment of 5 sentences using obtained model when source is German and target is English (Training sentences are not used)

1. Das digitale Fernsehen ist die einzige Konvergenztechnik, die keine allgemeine Interoperabilität aufweist.

   Best Alignment
   is is the the is the only interoperability. the without without interoperability. is

2. Zur Förderung der Interoperabilität zwischen den Netzen und Geräten für den Zugang zum digitalen Fernsehen und solchen für den Zugang zu interaktiven Diensten werden Vorschriften erlassen werden müssen.

   Best Alignment
   be to promote the services. between networks access and services. networks networks access access digital services. and required networks networks access access services. services. will required to will used

3. Eine solche Interoperabilität kann durch die Festlegung und Lizenzierung von Schlüsselstandards, Schnittstellen und Programmierwerkzeugen herbeigeführt werden, die zum Erreichen des Endnutzers erforderlich sind.

   Best Alignment
   be end end user. can through those standards, and user. of user. user. and user. be be those end can of user. necessary necessary

4. Dies fördert die Interoperabilität, ohne die Innovation zu untergraben.

   Best Alignment
   This This This will will without will innovation. promote innovation.

5. Dritten müssen diese Informationen zum selben Zeitpunkt zur Verfügung stehen wie den Diensten der Inhaber der Übergangsstellen.

   Best Alignment
   need Third need this information time same time information information at as access services. the same the services.

Alignment of 5 sentences using obtained model when source is English and target is German (Training sentences used)

1. Madam President, first of all I should like to thank Mr Koch for his report which has, at its heart, the issue of transport safety.

   Best Alignment
   Frau Präsidentin, Präsidentin, zunächst der zunächst ich Verkehrssicherheit möchte zum danken, Herrn Koch für seinen Bericht hat. Verkehrssicherheit Verkehrssicherheit seinen Verkehrssicherheit der Thema der Verkehrssicherheit Verkehrssicherheit

2. Indeed you responded in kind, as I knew you would do.

   Best Alignment
   Sie beantwortet, Sie beantwortet, so beantwortet, wie ich wußte, Sie würden. beantwortet,

3. We need to move that agenda forward with great urgency now.

   Best Alignment
   Hier dringender dringender jetzt dringender Handlungsbedarf. dringender dringender dringender dringender dringender dringender

4. Mr President, I too agree that the convention that we are examining is but a small step - and still too small a step - towards the construction of the common area of freedom, security and justice which Europe will hopefully one day become.

   Best Alignment
   noch Herr Präsident, ich verkörpern stimme daß den verkörpern daß uns uns verkörpern darstellt, kleiner ein kleiner Schritt - und noch verkörpern kleiner ein Schritt - Schritt den verkörpern des den gemeinsamen Rechts des Freiheit, Sicherheit und Freiheit, darstellt, Europa hoffentlich verkörpern Tages Tages verkörpern

5. So far, unfortunately, in Spain and France, just the opposite has been happening as regards the on-going conflict in the Basque country.

   Best Alignment
   Was passiert. passiert. leider in Spanien und Frankreich passiert. den Gegenteil Konflikt Konflikt passiert. so betrifft, den passiert. Konflikt in den Baskenland Baskenland

Alignment of 5 sentences using obtained model when source is English and target is German (Training sentences are not used)

1. Digital TV is the only convergent technology without basic interoperability.

   Best Alignment
   Das aufweist. ist ist die einzige aufweist. die allgemeine die aufweist.

2. Regulatory measures will be required to promote interoperability between the networks and devices that will be used to access digital TV and interactive services.

   Best Alignment
   Zur müssen. Förderung werden werden solchen zu Förderung müssen. zwischen den Förderung und müssen. Fernsehen werden werden werden zu Zugang digitalen zum und müssen. Diensten

3. Interoperability can be achieved through declaration and licensing of those key standards, interfaces and authoring tools necessary to reach the end user.

   Best Alignment
   Eine sind. kann werden, durch durch solche und werden, des werden, Eine Festlegung sind. und sind. und erforderlich Erreichen Eine des Eine sind.

4. This will promote interoperability without undermining innovation.

   Best Alignment
   Dies Dies die zu untergraben. ohne zu Innovation

5. Third parties will need access to this information at the same time as the gateway owners' own services.

   Best Alignment
   stehen Dritten stehen Verfügung müssen Informationen zur diese Informationen Zeitpunkt den selben Zeitpunkt wie den Übergangsstellen. Übergangsstellen. wie Diensten

The error generated by IBM Model 1 is substantial since it does not involve Language model input as checkpoint for sentences. This can be reduced by feeding more input sentences.
We can also calculate the error percentage by comparing the translated sentences to the actual Language model.

Submitted by-
BRIJ MOHAN LAL SRIVASTAVA
201307694
brijmohanlal.s@research.iiit.ac.in