LING 575 2011 Statistical Machine Translation

Homework 2 Solutions

Out April 5, due April 20 by midnight PST

100 points

The mean for this assignment is 68.25.

Problem 1. IBM word alignment models. [25 points]

For this problem we will use the following notation. ***f*** denotes a source sentence of length l_f and ***e*** denotes a corresponding target translation of length l_e. We can denote the words in the two sentences by $\boldsymbol{f} = [f_1, f_2, .., f_{l_f}]$ and $\mathbf{e} = [e_1, e_2, \ldots e_{l_e}]$. The lexical translation probabilities are written as $t(e|f)$ for all word pairs and $t(e|NULL)$ for inserted target words.

a. According to IBM models 1 and 2, how many alignments are possible between 3 target words and 5 source words? [5 points]

Solution: Each target word can align to one of the 5 source words or NULL, so we have 6 options per target word.

6^3 = 6*6*6 = 216

b. Write an expression for the posterior probability that all target words align to NULL, given the two sentences, according to IBM Model 1. To state this more formally denote by ***A*** the event that all target words align to NULL. We are looking for the probability $P(A|\boldsymbol{e}, \boldsymbol{f})$ in terms of the model parameters and the words in the sentences. [10 points]

$$\prod_{j=1}^{l_e} \frac{\boldsymbol{t(e_j|NULL)}}{\boldsymbol{t(e_j|NULL)} + \sum_{i=1}^{l_f} \boldsymbol{t(e_j|f_i)}}$$

c. Same as b) but using IBM Model 3: Write an expression for the posterior probability that all target words align to NULL, given the two sentences, according to IBM Model 3. To state this more formally denote by ***A*** the event that all target words align to NULL. We are looking for the

probability $P(A|\boldsymbol{e},\boldsymbol{f})$ in terms of the model parameters and the words in the sentences. [5 points]

Solution: According to IBM Model 3, each source word first generates several target words, and then each of the generated target words can give rise to an inserted (null-generated) word with some probability. If there are no target words generated in the first step, there can be no inserted target words.

Therefore it is not possible for all target words to align to NULL (the probability is zero).

d. Explain why the intersection of IBM model alignments in two directions (one using ***f*** as source and ***e*** as target, and one using ***e*** as source and ***f*** as target) is at most one to one***.*** At most one to one means that for each position *i* in ***f*** there exists at most one position *j* in ***e*** such that *i* is aligned to *j*, and that for each position *j* in ***e*** there is at most one position *i* in ***f*** such that *i* is aligned to *j*. [5 points]

Solution: We will prove this by contradiction. Suppose there is a pair of aligned words (f_i, e_j) and another pair (f_i, e_j') where *j* is not equal to *j'*. Since these alignment links are in the intersection of alignments in both directions, they exist in the alignments proposed by the two directions. In particular they should both belong to the alignments when ***f*** is target and ***e*** is source. But in this direction it is not possible for f_i to align to two distinct source words. Therefore, the alignments are at most one to one.

Problem 2. Implementing word-alignment models. [55 points]

Implement IBM Models 1 and 2 in a high-level programming language. You can use the pseudo-code in the textbook for guidance.

You need to create data structures for storing model parameters and to estimate the parameters from parallel sentences.

A 10,000 sentence corpus is located at **/dropbox/10-11/575SMT/HW2/data10kenes** on **patas.ling.washington.edu**. The 12-sentence toy English/Spanish corpus from the beginning of Lecture 2 is at **/dropbox/10-11/575SMT/HW2/toy**

a. Implement IBM Model 1 and train it on the union of data10kenes and toy, using 5 EM iterations. **[25 points]**
    i. Report the training data log-likelihood after each EM iteration.

ii. Generate the most likely English-to-Spanish and Spanish-to-English alignments on the toy corpus. What errors, if any, has the model made for each direction?

Solution: In this table we report perplexity instead of log-likelihood (perplexity = -log-likelihood)

$$-\sum_i \log_2 p(e_i|f_i)$$

	Direction source=English target=Spanish	**Direction source=Spanish target=English**
Perplexity iteration 1	**3577619.71**	**3094425.43**
Perplexity iteration 2	**1726944.56**	**1519413.59**
Perplexity iteration 3	**1511763.46**	**1321281.88**
Perplexity iteration 4	**1423673.64**	**1230832.80**
Perplexity iteration 5	**1391477.87**	**1196167.59**

Perplexity should go down with training iterations and perplexity when English is the target language should be lower. Since English is less morphologically rich, this makes sense.

Model 1 can make multiple errors on the example sentences, often aligning multiple target words to the same source word, or preferring to align to a less frequent source word. The list of alignments is at the end of this document.

Grading: 25 points if everything looks correct, with partial credit levels of 20, 15, 10, or 5 points depending on how different results were obtained.

b. Generalize your implementation to train IBM model 2 on the same data, using 5 EM iterations and starting from the learned parameters for IBM Model 1. **[30 points]**
    i. Report the training data log-likelihood after each EM iteration.
    ii. Generate the most likely English-to-Spanish and Spanish-to-English alignments on the toy corpus. What errors, if any, has the model made for each direction?

Optional: If you would like to try out a different parameterization of the distortion probabilities in IBM Model 2, you can introduce parameters depending on the difference between target position *j* and source position *i*.

Then we would have parameters like $\gamma(i-j)$, which might range from e.g. $\gamma(-20)$ to $\gamma(+20)$, where the range is determined by the largest possible difference according to the training data. It is also good

to introduce a special parameter for alignment to NULL, instead of treating NULL as a word at position 0 and looking up $\gamma(0-j)$. Then we would have the distortion parameters $\gamma(null), \gamma(-20), \ldots\ \gamma(0), \ldots \gamma(+20)$.

To obtain the normalized distortion probabilities (denoted by $a(i|j, l_e, l_f)$ in line 36 of the pseudocode for model 2 in Figure 4.7 of the textbook (page 99)), we would just normalize the gamma parameters appropriately:

$$a(i|j, l_e, l_f) = \frac{\gamma(i-j)}{\gamma(null) + \sum_{i'=1}^{l_f} \gamma(i'-j)}$$

If *i* is 0 in the above equation we use $\gamma(null)$ in the numerator.

This M-step is not exact and the exact one would be more complicated but this is a good approximation.

We then zero out the γ parameters for subsequent iterations in line 7 of the pseudo-code and increment them in line 24.

Solution: In this table we report perplexity instead of log-likelihood (perplexity = -log-likelihood). The measures are for a model using differences between source and target positions, with range from -10 to +10 (all differences greater than 10 are capped to 10).

	Direction source=English target=Spanish	**Direction source=Spanish target=English**
Perplexity iteration 1	**1316081.11**	**1117516.58**
Perplexity iteration 2	**1214379.85**	**1011928.07**
Perplexity iteration 3	**1171663.82**	**973033.44**
Perplexity iteration 4	**1147981.62**	**952722.39**
Perplexity iteration 5	**1133858.39**	**941093.05**

Similarly to model one perplexity we should observe decrease with iterations and lower perplexity when English is the target language. Additionally, Model 2 should reach lower perplexity than Model 1, since it adds parameters for estimating alignment probabilities whereas Model one uses uniform probabilities, which is a special case.

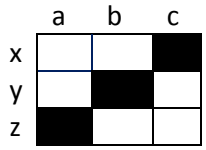
Generally the alignments obtained by Model 1 are better. The complete list of alignments is at the end of this document.

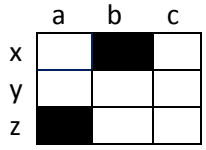
Grading: 30 points if everything looks correct, with partial credit possible of 25, 20,15,10, or 5 points.

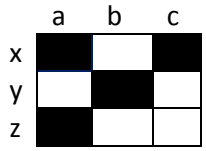
Problem 3. Phrase extraction from aligned sentence pairs [20 points]

For this problem, we will use the definition of the phrase extraction heuristic defined in Lecture 3 and in the textbook (definition 5.3 on page 131).

1. Consider the following three word alignment examples. For each, how many and which phrase-pairs can be extracted? [10 points]





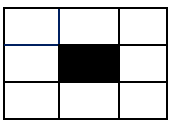


Solution:

1 [x,c] [y,b] [z,a] [xy,bc] [yz,bc][xyz,abc]
2 [x,b] [z,a] [x,bc][xy,b][xy,bc][yz,a][xyz,ab][xyz,abc]
3 [y,b] [xyz,abc]

2. What is the maximum number of phrase pairs that can be extracted from a three-by-three word example, if you are free to choose the word-alignment to maximize the number of extracted phrase pairs? Show such an alignment and report the number of phrase-pairs extracted. [10 points]

16 phrase-pairs from this alignment:



Alignments from Model 1 when English is source and Spanish is target

no comment

garcia y asociados .

NULL ({ / / }) garcia ({ 1 / / }) and ({ 2 / / }) associates ({ 3 / / }) . ({ 4 / / })

no comment

carlos garcia tiene tres asociados .

NULL ({ / / }) carlos ({ 1 / / }) garcia ({ 2 3 / / }) has ({ / / }) three ({ 4 / / }) associates ({ 5 / / }) . ({ 6 / / })

no comment

sus asociados no son fuertes .

NULL ({ / / }) his ({ / / }) associates ({ 1 2 / / }) are ({ 4 / / }) not ({ 3 / / }) strong ({ 5 / / }) . ({ 6 / / })

no comment

garcia tambien tiene una empresa .

NULL ({ / / }) garcia ({ 1 2 3 / / }) has ({ / / }) a ({ 4 / / }) company ({ 5 / / }) also ({ / / }) . ({ 6 / / })

no comment

sus clientes estan enfadados .

NULL ({ / / }) its ({ 1 / / }) clients ({ 2 / / }) are ({ / / }) angry ({ 3 4 / / }) . ({ 5 / / })

no comment

los asociados tambien estan enfadados .

NULL ({ / / }) the ({ / / }) associates ({ 2 / / }) are ({ 1 / / }) also ({ / / }) angry ({ 3 4 5 / / }) . ({ 6 / / })

no comment

los clients y los asociados son enemigos .

NULL ({ / / }) the ({ / / }) clients ({ 2 / / }) and ({ 3 / / }) the ({ / / }) associates ({ 5 / / }) are ({ 1 4 6 / / }) enemies ({ 7 / / }) . ({ 8 / / })

no comment

la empresa tiene tres grupos .

NULL ({ / / }) the ({ 1 / / }) company ({ 2 / / }) has ({ 3 / / }) three ({ 4 / / }) groups ({ 5 / / }) . ({ 6 / / })

no comment

sus grupos estan en europa .

NULL ({ / / }) its ({ 1 / / }) groups ({ 2 3 / / }) are ({ / / }) in ({ 4 / / }) europe ({ 5 / / }) . ({ 6 / / })

no comment

los grupos modernos venden medicinas fuertes .

NULL ({ / / }) the ({ / / }) modern ({ / / }) groups ({ 1 2 / / }) sell ({ 4 / / }) strong ({ / / }) pharmaceuticals ({ 3 5 6 / / }) . ({ 7 / / })

no comment

los grupos no venden zanzanina .

NULL ({ / / }) the ({ / / }) groups ({ 2 / / }) do ({ / / }) not ({ 3 / / }) sell ({ / / }) zenzanine ({ 1 4 5 / / }) . ({ 6 / / })

no comment

los grupos pequenos no son modernos .

NULL ({ / / }) the ({ / / }) small ({ / / }) groups ({ 2 / / }) are ({ 1 5 / / }) not ({ 4 / / }) modern ({ 3 6 / / }) . ({ 7 / / })

Model 1 alignment when English is target and Spanish is source

no comment

garcia and associates .

NULL ({ / / }) garcia ({ 1 / / }) y ({ 2 / / }) asociados ({ 3 / / }) . ({ 4 / / })

no comment

carlos garcia has three associates .

NULL ({ / / }) carlos ({ 1 / / }) garcia ({ 2 / / }) tiene ({ 3 / / }) tres ({ 4 / / }) asociados ({ 5 / / }) . ({ 6 / / })

no comment

his associates are not strong .

NULL ({ / / }) sus ({ / / }) asociados ({ 1 2 / / }) no ({ 4 / / }) son ({ 3 / / }) fuertes ({ 5 / / }) . ({ 6 / / })

no comment

garcia has a company also .

NULL ({ / / }) garcia ({ 1 / / }) tambien ({ 5 / / }) tiene ({ 2 / / }) una ({ 3 / / }) empresa ({ 4 / / }) . ({ 6 / / })

no comment

its clients are angry .

NULL ({ / / }) sus ({ / / }) clientes ({ 2 / / }) estan ({ 1 3 / / }) enfadados ({ 4 / / }) . ({ 5 / / })

no comment

the associates are also angry .

NULL ({ / / }) los ({ 1 / / }) asociados ({ 2 / / }) tambien ({ 4 / / }) estan ({ 3 / / }) enfadados ({ 5 / / }) . ({ 6 / / })

no comment

the clients and the associates are enemies .

NULL ({ / / }) los ({ 1 4 / / }) clients ({ 2 / / }) y ({ 3 / / }) los ({ / / }) asociados ({ 5 / / }) son ({ 6 / / }) enemigos ({ 7 / / }) . ({ 8 / / })

no comment

the company has three groups .

NULL ({ / / }) la ({ 1 / / }) empresa ({ 2 / / }) tiene ({ 3 / / }) tres ({ 4 / / }) grupos ({ 5 / / }) . ({ 6 / / })

no comment

its groups are in europe .

NULL ({ / / }) sus ({ / / }) grupos ({ 2 / / }) estan ({ 1 3 / / }) en ({ 4 / / }) europa ({ 5 / / }) . ({ 6 / / })

no comment

the modern groups sell strong pharmaceuticals .

NULL ({ / / }) los ({ 1 / / }) grupos ({ 3 / / }) modernos ({ 2 / / }) venden ({ 4 / / }) medicinas ({ 6 / / }) fuertes ({ 5 / / }) . ({ 7 / / })

no comment

the groups do not sell zenzanine .

NULL ({ / / }) los ({ 1 / / }) grupos ({ 2 / / }) no ({ 4 / / }) venden ({ 5 / / }) zanzanina ({ 3 6 / / }) . ({ 7 / / })

no comment

the small groups are not modern .

NULL ({ / / }) los ({ 1 / / }) grupos ({ 3 / / }) pequenos ({ 2 / / }) no ({ 5 / / }) son ({ 4 / / }) modernos ({ 6 / / }) . ({ 7 / / })

Model 2 alignments when Spanish is source and English is target

no comment

garcia and associates .

NULL ({ / / }) garcia ({ 1 / / }) y ({ 2 / / }) asociados ({ 3 / / }) . ({ 4 / / })

no comment

carlos garcia has three associates .

NULL ({ / / }) carlos ({ 1 / / }) garcia ({ 2 / / }) tiene ({ 3 / / }) tres ({ 4 / / }) asociados ({ 5 / / }) . ({ 6 / / })

no comment

his associates are not strong .

NULL ({ / / }) sus ({ / / }) asociados ({ 1 2 / / }) no ({ 4 / / }) son ({ 3 / / }) fuertes ({ 5 / / }) . ({ 6 / / })

no comment

garcia has a company also .

NULL ({ / / }) garcia ({ 1 / / }) tambien ({ 5 / / }) tiene ({ 2 / / }) una ({ 3 / / }) empresa ({ 4 / / }) . ({ 6 / / })

no comment

its clients are angry .

NULL ({ / / }) sus ({ 1 / / }) clientes ({ 2 / / }) estan ({ 3 / / }) enfadados ({ 4 / / }) . ({ 5 / / })

no comment

the associates are also angry .

NULL ({ / / }) los ({ 1 / / }) asociados ({ 2 / / }) tambien ({ 4 / / }) estan ({ 3 / / }) enfadados ({ 5 / / }) . ({ 6 / / })

no comment

the clients and the associates are enemies .

NULL ({ / / }) los ({ 1 / / }) clients ({ 2 / / }) y ({ 3 / / }) los ({ 4 / / }) asociados ({ 5 / / }) son ({ 6 / / }) enemigos ({ 7 / / }) . ({ 8 / / })

no comment

the company has three groups .

NULL ({ / / }) la ({ 1 / / }) empresa ({ 2 / / }) tiene ({ 3 / / }) tres ({ 4 / / }) grupos ({ 5 / / }) . ({ 6 / / })

no comment

its groups are in europe .

NULL ({ / / }) sus ({ 1 / / }) grupos ({ 2 / / }) estan ({ 3 / / }) en ({ 4 / / }) europa ({ 5 / / }) . ({ 6 / / })

no comment

the modern groups sell strong pharmaceuticals .

NULL ({ / / }) los ({ 1 / / }) grupos ({ 3 / / }) modernos ({ 2 / / }) venden ({ 4 / / }) medicinas ({ 6 / / }) fuertes ({ 5 / / }) . ({ 7 / / })

no comment

the groups do not sell zenzanine .

NULL ({ / / }) los ({ 1 / / }) grupos ({ 2 / / }) no ({ 4 / / }) venden ({ 5 / / }) zanzanina ({ 3 6 / / }) . ({ 7 / / })

no comment

the small groups are not modern .

NULL ({ / / }) los ({ 1 / / }) grupos ({ 3 / / }) pequenos ({ 2 / / }) no ({ 5 / / }) son ({ 4 / / }) modernos ({ 6 / / }) . ({ 7 / / })

Model 2 alignments when English is source and Spanish is target

no comment

garcia y asociados .

NULL ({ / / }) garcia ({ 1 / / }) and ({ 2 / / }) associates ({ 3 / / }) . ({ 4 / / })

no comment

carlos garcia tiene tres asociados .

NULL ({ / / }) carlos ({ 1 / / }) garcia ({ 2 3 / / }) has ({ / / }) three ({ 4 / / }) associates ({ 5 / / }) . ({ 6 / / })

no comment

sus asociados no son fuertes .

NULL ({ / / }) his ({ 1 / / }) associates ({ 2 / / }) are ({ 4 / / }) not ({ 3 / / }) strong ({ 5 / / }) . ({ 6 / / })

no comment

garcia tambien tiene una empresa .

NULL ({ / / }) garcia ({ 1 2 3 / / }) has ({ / / }) a ({ 4 / / }) company ({ 5 / / }) also ({ / / }) . ({ 6 / / })

no comment

sus clientes estan enfadados .

NULL ({ / / }) its ({ 1 / / }) clients ({ 2 / / }) are ({ / / }) angry ({ 3 4 / / }) . ({ 5 / / })

no comment

los asociados tambien estan enfadados .

NULL ({ / / }) the ({ 1 / / }) associates ({ 2 3 / / }) are ({ / / }) also ({ / / }) angry ({ 4 5 / / }) . ({ 6 / / })

no comment

los clients y los asociados son enemigos .

NULL ({ / / }) the ({ 1 / / }) clients ({ 2 / / }) and ({ 3 / / }) the ({ 4 / / }) associates ({ 5 / / }) are ({ 6 / / }) enemies ({ 7 / / }) . ({ 8 / / })

no comment

la empresa tiene tres grupos .

NULL ({ / / }) the ({ 1 / / }) company ({ 2 / / }) has ({ 3 / / }) three ({ 4 / / }) groups ({ 5 / / }) . ({ 6 / / })

no comment

sus grupos estan en europa .

NULL ({ / / }) its ({ 1 / / }) groups ({ 2 3 / / }) are ({ / / }) in ({ 4 / / }) europe ({ 5 / / }) . ({ 6 / / })

no comment

los grupos modernos venden medicinas fuertes .

NULL ({ / / }) the ({ 1 / / }) modern ({ 3 / / }) groups ({ 2 / / }) sell ({ 4 / / }) strong ({ / / }) pharmaceuticals ({ 5 6 / / }) . ({ 7 / / })

no comment

los grupos no venden zanzanina .

NULL ({ / / }) the ({ 1 / / }) groups ({ 2 / / }) do ({ / / }) not ({ 3 / / }) sell ({ 4 / / }) zenzanine ({ 5 / / }) . ({ 6 / / })

no comment

los grupos pequenos no son modernos .

NULL ({ / / }) the ({ 1 / / }) small ({ 3 / / }) groups ({ 2 / / }) are ({ 5 / / }) not ({ 4 / / }) modern ({ 6 / / }) . ({ 7 / / })