**CMPE 590 SP.TOP. IN CMPE: MACHINE TRANSLATION**

**MIDTERM ANSWERS**

1. Using a controlled language in machine translation indicates that the complexity of sentences is decreased and the sentences are written in a more standard form. In the context of machine translation, controlled language is used to decrease the ambiguity in the text and thus to improve the translation quality.
2. The text should be divided into separate and shorter sentences. Complicated grammatical constructions should be avoided. Ambiguous words and idiomatic expressions should be avoided as much as possible.
3. The initial t(e|f) distribution:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |
| book | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |
| open | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |
| door | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |
| window | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 |

1st iteration:

1st sentence:

s-total(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | wonderful | book | open | door | window |
|  | 0.40 | 0.40 |  |  |  |

count(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.50 | 0.50 |  |  |  |
| book | 0.50 | 0.50 |  |  |  |
| open |  |  |  |  |  |
| door |  |  |  |  |  |
| window |  |  |  |  |  |
| *total(f)* | 1.00 | 1.00 |  |  |  |

2nd sentence:

s-total(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | wonderful | book | open | door | window |
|  |  | 0.40 | 0.40 |  |  |

count(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.50 | 0.50 |  |  |  |
| book | 0.50 | 1.00 | 0.50 |  |  |
| open |  | 0.50 | 0.50 |  |  |
| door |  |  |  |  |  |
| window |  |  |  |  |  |
| *total(f)* | 1.00 | 2.00 | 1.00 |  |  |

3rd sentence:

s-total(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | wonderful | book | open | door | window |
|  |  |  | 0.60 | 0.60 | 0.60 |

count(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.50 | 0.50 |  |  |  |
| book | 0.50 | 1.00 | 0.50 |  |  |
| open |  | 0.50 | 0.83 | 0.33 | 0.33 |
| door |  |  | 0.33 | 0.33 | 0.33 |
| window |  |  | 0.33 | 0.33 | 0.33 |
| *total(f)* | 1.00 | 2.00 | 2.00 | 1.00 | 1.00 |

Then, t(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.50 | 0.25 |  |  |  |
| book | 0.50 | 0.50 | 0.25 |  |  |
| open |  | 0.25 | 0.42 | 0.33 | 0.33 |
| door |  |  | 0.17 | 0.33 | 0.33 |
| window |  |  | 0.17 | 0.33 | 0.33 |

2nd iteration:

1st sentence:

s-total(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | wonderful | book | open | door | window |
|  | 0.75 | 1.00 |  |  |  |

count(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.66 | 0.25 |  |  |  |
| book | 0.66 | 0.50 |  |  |  |
| open |  |  |  |  |  |
| door |  |  |  |  |  |
| window |  |  |  |  |  |
| *total(f)* | 1.32 | 0.75 |  |  |  |

2nd sentence:

s-total(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | wonderful | book | open | door | window |
|  |  | 0.75 | 0.67 |  |  |

count(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.66 | 0.25 |  |  |  |
| book | 0.66 | 1.16 | 0.33 |  |  |
| open |  | 0.37 | 0.63 |  |  |
| door |  |  |  |  |  |
| window |  |  |  |  |  |
| *total(f)* | 1.32 | 1.78 | 0.96 |  |  |

3rd sentence:

s-total(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | wonderful | book | open | door | window |
|  |  |  | 1.08 | 0.83 | 0.83 |

count(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.66 | 0.25 |  |  |  |
| book | 0.66 | 1.16 | 0.33 |  |  |
| open |  | 0.37 | 1.00 | 0.31 | 0.31 |
| door |  |  | 0.20 | 0.40 | 0.40 |
| window |  |  | 0.20 | 0.40 | 0.40 |
| *total(f)* | 1.32 | 1.78 | 1.73 | 1.11 | 1.11 |

Then, t(e|f)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | harika | kitap | açık | kapı | penceresi |
| wonderful | 0.50 | 0.14 |  |  |  |
| book | 0.50 | 0.65 | 0.19 |  |  |
| open |  | 0.21 | 0.58 | 0.28 | 0.28 |
| door |  |  | 0.12 | 0.36 | 0.36 |
| window |  |  | 0.12 | 0.36 | 0.36 |

1. The combined word alignment table is as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Kafka | en | iyi | yazarlardan | biridir |
| Kafka | X |  |  |  |  |
| is |  |  |  |  | X |
| one |  |  |  |  | X |
| of |  |  |  | X |  |
| the |  |  |  | X |  |
| best |  |  | X |  |  |
| authors |  |  |  | X |  |

The algorithm first considers the word alignments that occur in both directions and places them in to the table (black marks).

1st step: For each square marked, the neighbors are checked. For a neighbor (i,j), if there is an alignment in one of the unidirectional tables and if at least one of the corresponding two words is still unaligned, we mark that square (red marks).

2nd step: We check each square (i,j). If there is an alignment in the English-Turkish table and if at least one of the corresponding two words is still unaligned, we mark that square (no such cases above).

3rd step: We check each square (i,j). If there is an alignment in the Turkish-English table and if at least one of the corresponding two words is still unaligned, we mark that square (no such cases above).

The extracted phrases are as follows:

|  |  |
| --- | --- |
| **English phrase** | **Turkish phrase** |
| Kafka | KafkaKafka en |
| Kafka is one of the best authors | Kafka en iyi yazarlardan biridir |
| is one | biridir |
| is one of the best authors | iyi yazarlardan biridiren iyi yazarlardan biridir |
| of the best authors | iyi yazarlardanen iyi yazarlardan |
| best | iyien iyi |

The algorithm scans all possible sequences of English words. For a sequence ei…ej, we determine the corresponding boundary fk…fl. If there are no marks above and below the rectangle ei…ej – fk…fl, we take it as a phrase. We also expand the rectangle to the left and right. If the same condition is met, we take that one as a phrase too.

1. The general formula is as follows, where f and e denote the input/output senteces, xi denotes the ith word, and $\overbar{x\_{i}}$ denotes the ith phrase. I is the number of phrases. starti and endi are related to the distance-based reordering model.

$$P\left(f\right)=\prod\_{i=1}^{I}∅\left(\overbar{f\_{i}}|\overbar{e\_{i}}\right)\*\prod\_{i=1}^{I}d\left(start\_{i}|end\_{i-1}-1\right)\*\prod\_{i=1}^{l\_{e}}p\_{LM}\left(e\_{i}|e\_{1}..e\_{i-1}\right)$$

Below are phrase translation tables for English phrases that are used in the example translations. The data are imaginary.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| I | go | to school | school | go to | fit | to college |
| ben | 0.90 | git | 0.20 | okula | 0.80 | okul | 0.70 | ben giderim | 0.10 | uyar | 0.40 | okula | 0.10 |
| bana | 0.10 | gitmek | 0.10 | okul | 0.05 | okula | 0.20 | gidiyorum | 0.30 | gidiyorum | 0.10 | bu okul | 0.10 |
|  |  | gidiyorum | 0.20 | okulda | 0.15 | okullar | 0.10 | oraya gitmek | 0.60 | uydu | 0.50 | koleje | 0.80 |
|  |  | gidiyorsun | 0.10 |  |  |  |  |  |  |  |  |  |  |
|  |  | gider | 0.40 |  |  |  |  |  |  |  |  |  |  |

Reordering model: We use a fixed distribution $d\left(x\right)=α^{\left|x\right|}$ with α=0.90.

Language model: Suppose that we use bigram language model. The probabilities for the bigrams in the examples are as follows.

|  |  |
| --- | --- |
| bigram | probability |
| <s> I | 0.20 |
| I go | 0.30 |
| go to | 0.50 |
| to school  | 0.10 |
| <s> school | 0.10 |
| school I | 0.10 |
| I fit | 0.20 |
| fit to | 0.30 |
| to college | 0.10 |

1st translation:

f : *Ben okula gidiyorum*

e: *I go to school*

$P\left(f\right)=\left(0.90\*0.20\*0.80\right)\*\left(0.90^{0}\*0.90^{1}\*0.90^{2}\right)\*\left(0.20\*0.30\*0.50\*0.10\right)=0.000315$

2nd translation:

f : *Ben okula gidiyorum*

e: *School I go to*

$P\left(f\right)=\left(0.20\*0.90\*0.30\right)\*\left(0.90^{1}\*0.90^{2}\*0.90^{1}\right)\*\left(0.10\*0.10\*0.30\*0.50\right)=0.000053$

3rd translation:

f : *Ben okula gidiyorum*

e: *I fit to college*

$$P\left(f\right)=\left(0.90\*0.10\*0.10\right)\*\left(0.90^{0}\*0.90^{1}\*0.90^{2}\right)\*\left(0.20\*0.20\*0.30\*0.10\right)=0.000008$$

Thus, the best translation is *I go to school*.