Assignment 4
Vijayalakshmi Dhamodharan

This document has the detailed analysis and the results for the Sentence compression project. The first section has the precision, recall, f-measure and the second section has the analysis and the comparison. This project has been implemented in Java and I have used ‘lpsolve ‘ for the ILP part. And I have used your scoring program for printing the scores.

Compression system 1 (no syntax):

Here words are only selected based on the weights which has been calculated by *tf-idf(word, doc) - .4 * depth(word) - .5*. Hence all the negative weights will be dropped even if it contributes to the syntax of the sentence. I have pasted few examples below

Orig: The Choice of the capital of Cornwall 's one-time china clay industry as the venue for the Liberal Democrats penultimate and most glitzy election rally last night was no accident
Human: The Choice of the capital of Cornwall 's one-time industry for the Liberal Democrats glitzy rally was no accident
Ours: Choice capital Cornwall one-time china clay venue Liberal Democrats penultimate rally no accident

Orig: Mr Ashdown descended on the area yesterday for the third time in his marathon campaign amid growing recognition among Tories that the South-West could prove their Achilles heel a danger underlined by the Press Association opinion poll which showed a swing against the Conservatives in the region of six per cent
Human: Mr Ashdown descended on the area for the third time amid recognition among Tories that the South-West could prove their Achilles heel a danger underlined by the poll which showed a swing against the Conservatives of six per cent

Ours: Mr Ashdown descended area third marathon amid growing recognition Tories South-West prove Achilles heel danger underlined Press Association opinion poll showed swing against Conservatives region

In the above examples the words like “the”, “of”, “for”, “was”, “on” has been dropped because of their low weights which has been calculated by their tf-idf scores.

The scores I got for the compression system 1 is as follows:

40283 compressed by humans to 27480 0.317826378373
40283 compressed by us to 22999 0.429064369585
Matched 16511 of 27480 human words
Printed 22999 words
Prec 0.717900778295 rec 0.600836972344 F 0.654173022445

The recall and f-measure is slightly greater than the values that you have provided.

Compression system 2 (a little syntax)

Here an additional constraint of “ adding a head word if a word has been selected” is added in addition to the no syntax one in order to inject some syntax to the sentence.

The scores are below:

40283 compressed by humans to 27480 0.317826378373
40283 compressed by us to 27966 0.30576173572
Matched 19876 of 27480 human words
Printed 27966 words
Prec 0.710720160195 rec 0.723289665211 F 0.716949825055

When we compare the results with the compression system1, the compression value has been decreased, but we have achieved a better accuracy. Since the sentences are more syntactical than the previous system.

For example take the below example

Orig: Mr Ashdown descended on the area yesterday for the third time in his marathon campaign amid growing recognition among Tories that the South-West could prove their Achilles heel a danger underlined by the Press Association opinion poll which showed a swing against the Conservatives in the region of six per cent

Human: Mr Ashdown descended on the area for the third time amid recognition among Tories that the South-West could prove their Achilles heel a danger underlined by the poll which showed a swing against the Conservatives of six per cent

Ours: Mr Ashdown descended on the area third time marathon campaign amid growing recognition among Tories South-West prove their Achilles heel a danger underlined by the Press Association opinion poll which showed a swing against the Conservatives in region six per cent

In the above examples the words like “on”, “their” has been added because of the additional constraint of adding the head word.

However, some of the example like below need more grammatical structures

Orig: The Choice of the capital of Cornwall 's one-time china clay industry as the venue for the Liberal Democrats penultimate and most glitzy election rally last night was no accident
Human: The Choice of the capital of Cornwall 's one-time industry for the Liberal Democrats glitzy rally was no accident
Ours: The Choice the capital Cornwall 's china clay industry the venue Liberal Democrats penultimate and most glitzy election rally last night was no accident

Still words like ‘of’ has been dropped by the system.

Compression system 3 (a bit more syntax)

Here an additional constraint of “ if a word is included and it has a child connected by an arc with a label that ought to be kept, the child will also be included”.

The scores of the above compressed system is

40283 compressed by humans to 27480 0.317826378373
40283 compressed by us to 32052 0.204329369709
Matched 22852 of 27480 human words
Printed 32052 words
Prec 0.712966429552 rec 0.831586608443 F 0.767721561513

Here the compression value has been decreased, but the recall and the accuracy has been improved noticeably due to the injection of more syntax.

Below are few examples:
Orig: Cutbacks in local defence establishments is also a factor in some constituencies
Human: Cutbacks in local defence is a factor
Ours: Cutbacks establishments **is** also factor some constituencies

Orig: She was 67
Human: She was 67
Ours: She was 67

Orig: But once she met Watkins whom she married in 1950 her career was set following him around the world and rearing their six children often alone
Human: once she met Watkins her career was set following him around the world and rearing their six children often alone
Ours: once she met Watkins whom she married in 1950 her career was set following around world and rearing their six children often alone

The sentences look more syntactical because the system preserves the words like was, is, and, whom etc. which is due to the additional dependencies that has been added.