CS134 Final Project Report

Hongyuan Shen (Harry)

The neural network is built within python file $neural\_network.py$. It has 3 feature representations: using all words in Arg1, Arg2 and Connective; using only the last 3 words in Arg1 (default), the first 3 words in Arg2 and Connective ($l3f3$); using cross product of unigrams: Arg1 × Arg2 ($word\_pairs$).

The neural network has 3 layers, hidden layers $h1$, $h2$ and output layer out. Each hidden layer has 256 neurons and the neural network has a learning rate 0.1 with activation function $Softmax$.

The training process uses mini-batch strategy and each iteration the neural network will train on 2000 random relations in the training set, then update based on the dev set. When it is converged, the program will run evaluation on the test set and (optionally) output a .json file for further evaluation which uses $scorer.py$.

When using the default feature, we have the following result:

Evaluation for all discourse relations:

Sense classification	precision	recall	F1
*Micro-Average	0.2646	0.2630	0.2638
Comparison.Concession	0.3333	0.0093	0.0182
Comparison.Contrast	0.0713	0.5357	0.1258
Contingency.Cause.Reason	0.3333	0.2632	0.2941
Contingency.Cause.Result	0.3125	0.0980	0.1493
Contingency.Condition	0.2838	0.8077	0.4200
EntRel	0.2909	0.3200	0.3048
Expansion.Alternative	1.0000	0.0667	0.1250
Expansion.Conjunction	0.4769	0.2925	0.3626
Expansion.Instantiation	0.6667	0.0455	0.0851
Expansion.Restatement	0.1429	0.0132	0.0242
Temporal.Asynchronous.Precedence	0.3692	0.4800	0.4174
Temporal.Asynchronous.Succession	0.4316	0.6613	0.5223
Temporal.Synchrony	0.4000	0.2642	0.3182

Overall parser performance: Precision 0.2646 Recall 0.2630 F1 0.2638

Evaluation for explicit discourse relations only:

Sense classification	precision	recall	F1
*Micro-Average	0.4360	0.3921	0.4129
Comparison.Concession	1.0000	0.0130	0.0256
Comparison.Contrast	0.1143	0.6897	0.1961
Contingency.Cause.Reason	0.5455	0.5294	0.5373
Contingency.Cause.Result	0.4444	0.2857	0.3478
Contingency.Condition	0.4565	0.8077	0.5833
Expansion.Alternative	1.0000	0.1000	0.1818
Expansion.Conjunction	0.7708	0.3592	0.4901
Expansion.Instantiation	0.5000	0.3333	0.4000
Expansion.Restatement	0.0000	0.0000	0.0000
Temporal.Asynchronous.Precedence	0.5750	0.5750	0.5750
Temporal.Asynchronous.Succession	0.6119	0.6613	0.6357
Temporal.Synchrony	0.5600	0.2800	0.3733

Overall parser performance: Precision 0.4360 Recall 0.3921 F1 0.4129

Evaluation for non-explicit discourse relations only (Implicit, EntRel, AltLex):

Sense classification	precision	recall	F1
*Micro-Average	0.1686	0.1531	0.1605
Comparison.Concession	0.0000	0.0000	0.0000
Comparison.Contrast	0.0407	0.3704	0.0733
Contingency.Cause.Reason	0.0741	0.0476	0.0580
Contingency.Cause.Result	0.1429	0.0270	0.0455
EntRel	0.3832	0.3200	0.3488
Expansion.Alternative	1.0000	0.0000	0.0000
Expansion.Conjunction	0.1919	0.1696	0.1801
Expansion.Instantiation	1.0000	0.0244	0.0476
Expansion.Restatement	0.2222	0.0137	0.0258
Temporal.Asynchronous.Precedence	0.0400	0.1000	0.0571
Temporal.Synchrony	0.0000	0.0000	0.0000

Overall parser performance: Precision 0.1686 Recall 0.1531 F1 0.1605

We can see that the accuracy is not very satisfiable.

However, when using last-3-first-3 as feature, we have a much enhanced score:

Evaluation for all discourse relations:

Sense classification	precision	recall	F1
*Micro-Average	0.4542	0.4516	0.4529
Comparison.Concession	0.5926	0.1495	0.2388
Comparison.Contrast	0.2073	0.3091	0.2482
Contingency.Cause.Reason	0.4394	0.3867	0.4113
Contingency.Cause.Result	0.4444	0.3265	0.3765
Contingency.Condition	0.5319	0.9615	0.6849
EntRel	0.3121	0.6850	0.4288
Expansion.Alternative	0.3333	0.0667	0.1111
Expansion.Conjunction	0.6556	0.6130	0.6336
Expansion.Instantiation	0.6667	0.0909	0.1600
Expansion.Restatement	0.3333	0.0530	0.0914
Temporal.Asynchronous.Precedence	0.7353	0.5000	0.5952
Temporal.Asynchronous.Succession	0.7400	0.5968	0.6607
Temporal.Synchrony	0.3837	0.6346	0.4783

Overall parser performance: Precision 0.4542 Recall 0.4516 F1 0.4529

Evaluation for explicit discourse relations only:

Sense classification	precision	recall	F1
*Micro-Average	0.7110	0.6727	0.6913
Comparison.Concession	0.7500	0.1948	0.3093
Comparison.Contrast	0.2576	0.5862	0.3579
Contingency.Cause.Reason	0.6585	0.8182	0.7297
Contingency.Cause.Result	0.5000	0.8333	0.6250
Contingency.Condition	0.6579	0.9615	0.7812
Expansion.Alternative	1.0000	0.1000	0.1818
Expansion.Conjunction	0.8571	0.8571	0.8571
Expansion.Instantiation	0.5000	0.3333	0.4000
Expansion.Restatement	0.4286	0.6000	0.5000
Temporal.Asynchronous.Precedence	0.9259	0.6250	0.7463
Temporal.Asynchronous.Succession	0.9024	0.5968	0.7184
Temporal.Synchrony	0.6226	0.6735	0.6471

Overall parser performance: Precision 0.7110 Recall 0.6727 F1 0.6913

Evaluation for non-explicit discourse relations only (Implicit, EntRel, AltLex):

Sense classification	precision	recall	F1
*Micro-Average	0.2717	0.2634	0.2675
Comparison.Concession	0.1429	0.0333	0.0541
Comparison.Contrast	0.0000	0.0000	0.0000
Contingency.Cause.Reason	0.0800	0.0476	0.0597
Contingency.Cause.Result	0.3750	0.1622	0.2264
EntRel	0.3309	0.6850	0.4463
Expansion.Alternative	0.0000	0.0000	0.0000
Expansion.Conjunction	0.1957	0.1593	0.1756
Expansion.Instantiation	0.7500	0.0732	0.1333
Expansion.Restatement	0.2941	0.0342	0.0613
Temporal.Asynchronous.Precedence	0.0000	0.0000	0.0000
Temporal.Synchrony	0.0000	0.0000	0.0000

Overall parser performance: Precision 0.2717 Recall 0.2634 F1 0.2675

The accuracy on explicit relations reached 71%, and 27% on implicit relations. The reason why it has a higher score may because that it works similarly as the attention mechanism. The words closer to the connective usually have more hints on the sense type.

Next, when using last-3-first-3 and also word pairs as features, we have the following result:

Evaluation for all discourse relations:

Sense classification	precision	recall	F1
*Micro-Average	0.4433	0.4367	0.4400
Comparison.Concession	0.4179	0.2617	0.3218
Comparison.Contrast	0.2632	0.2727	0.2679
Contingency.Cause.Reason	0.4098	0.3425	0.3731
Contingency.Cause.Result	0.2778	0.2830	0.2804
Contingency.Condition	0.7097	0.8462	0.7719
EntRel	0.3142	0.7400	0.4411
Expansion.Alternative	0.2857	0.1333	0.1818
Expansion.Conjunction	0.7789	0.4874	0.5996
Expansion.Instantiation	0.3333	0.0455	0.0800
Expansion.Restatement	0.3611	0.1722	0.2332
Temporal.Asynchronous.Precedence	0.6579	0.5000	0.5682
Temporal.Asynchronous.Succession	0.6562	0.6562	0.6562
Temporal.Synchrony	0.3594	0.4340	0.3932

Overall parser performance: Precision 0.4433 Recall 0.4367 F1 0.4400

Evaluation for explicit discourse relations only:

Sense classification	precision	recall	F1
*Micro-Average	0.6801	0.6079	0.6420
Comparison.Concession	0.6429	0.3506	0.4538
Comparison.Contrast	0.3333	0.5172	0.4054
Contingency.Cause.Reason	0.5750	0.7419	0.6479
Contingency.Cause.Result	0.4000	0.8750	0.5490
Contingency.Condition	0.8148	0.8462	0.8302
Expansion.Alternative	0.3333	0.1000	0.1538
Expansion.Conjunction	0.9103	0.6927	0.7867
Expansion.Instantiation	0.5000	0.3333	0.4000
Expansion.Restatement	0.1667	0.6000	0.2609
Temporal.Asynchronous.Precedence	0.7576	0.6250	0.6849
Temporal.Asynchronous.Succession	0.8400	0.6562	0.7368
Temporal.Synchrony	0.5000	0.4600	0.4792

Overall parser performance: Precision 0.6801 Recall 0.6079 F1 0.6420

Evaluation for non-explicit discourse relations only (Implicit, EntRel, AltLex):

Sense classification	precision	recall	F1
*Micro-Average	0.3011	0.2910	0.2960
Comparison.Concession	0.0400	0.0333	0.0364
Comparison.Contrast	0.0000	0.0000	0.0000
Contingency.Cause.Reason	0.0952	0.0476	0.0635
Contingency.Cause.Result	0.0526	0.0270	0.0357
EntRel	0.3474	0.7400	0.4728
Expansion.Alternative	0.2500	0.2000	0.2222
Expansion.Conjunction	0.3023	0.1150	0.1667
Expansion.Instantiation	0.2500	0.0244	0.0444
Expansion.Restatement	0.4259	0.1575	0.2300
Temporal.Asynchronous.Precedence	0.0000	0.0000	0.0000
Temporal.Synchrony	0.0000	0.0000	0.0000

Overall parser performance: Precision 0.3011 Recall 0.2910 F1 0.2960

We can see that by adding word pairs we get a slightly better score on implicit relations.