Supplemental Material

1. Background and Motivation

The examples below show instances where identifying discourse relations are very helpful in information retrieval and extraction tasks. The connective *also* in Example 1 suggests that its sentence, taken in isolation, does not provide the complete information about where the IgMhi cells are found. In order to complete this information, the previous sentence must be taken into account, as well.

Example 1: *In control B6 mice, IgMhi cells (in red) were present in the MZ, outside of the MOMA-1+ cells.* IgMhi MZ B cells were also found outside of the MOMA-1 ring in p50-/- mice, but in reduced numbers, as expected from the drastic reductions in MZ B cells in these mice. (Conjunction)

Discourse relations can also be useful for categorizing citations and the relations between the citations to enhance information retrieval: the connective *In contrast* in Example 2 signals a contrast relation between two cited articles, "48" and "49," mentioned in two different sentences.

Example 2**:** *The importance of PU.1 for Btk gene regulation is underlined by the fact that the absence of PU.1 leads to a two- to 3-fold reduction of Btk expression (48).* **In contrast**, the deficiency of Sp1 that also stimulates Btk promoter activity together with PU.1 (49) had no influence on Btk expression (49). (Contrast)

For summarization tasks, it is useful to identify summary sentences, as well as the larger text segments that such sentences summarize. Connectives like *In conclusion* in Example 3 are important indicators of such relations.

Example 3**:** *Consistent with our binding studies, we observed that BOB.1/OBF.1 together with Oct2 was able to activate the murine Btk promoter <150-fold in a dose-dependent manner (Figure 5A, B and data not shown). Transfection experiments using NIH/3T3 cells revealed that BOB.1/OBF.1 together with PU.1 only marginally enhanced PU.1-mediated Btk promoter activity. In contrast, co-transfection of Oct2 together with PU.1 stimulated PU.1-mediated Btk promoter activity significantly (from 6- to 75-fold). Moreover, co-transfection of PU.1 together with Oct2 and BOB.1/OBF.1 led to an even*

stronger and synergistic activation (325-fold) of the murine Btk promoter (Figure 5C). **In conclusion**, these findings indicate that the transcriptional coactivator BOB.1/OBF.1 regulates the Btk promoter activity in B cells in vitro as well as in vivo, in concert with Oct and PU.1 proteins. (Restatement: Generalization)

Causal and justification relations also constitute a very important part of the knowledge dealt within information extraction: for example, the connective *since* in Example 4 signals a causal relation between the two clauses. In other words, the fact that "HeLa cells do not express Oct2" is the reason (or reason for believing) that "the addition of an anti-Oct2 antibody did not interfere with complex formation."

***Example 4*:** *The addition of an anti-Oct2 antibody did not interfere with complex formation,* **since** HeLa cells do not express Oct2. (Cause: Reason)

Example 5 illustrates the importance of accurately disambiguating ADE causal relations. Here, with a co-occurrence approach, both "Solu-Medro" and "cyclosporin" present themselves as the causes of the "acute renal failure." On the other hand, by recognizing the connective *so* and its arguments, we can accurately select "cyclosporin" as the drug causing the renal failure.

***Example 5*:** In the emergency department, he was given one dose of Solu-Medro 500 mg, however, he was found to have elevated cyclosporin levels at 679, **so** this was thought to be the likely cause of his acute renal failure and his cyclosporin was temporarily held. Since that time, on his hospital day #1, his cyclosporin levels trended down to the point at which there were just slightly over 100 on hospital day #3 and cyclosporin was reinitiated at lower doses. He was dialyzed on admission with removal of 4 liters of fluid, CVM, BK, and LDH were sent from dialysis. His creatinine improved, so further dialysis and biopsy were deemed unnecessary. (A narrative excerpt released by the i2b2 organizer [44].)

Discourse connective identification is a hard task. Compare, for instance, the use of the word "briefly" in the (a) and (b) sentences in Example 6. In Example 6a, the word "Briefly" is used to express the elaboration or specification relation in discourse, whereas the same word in Example 6b functions as a temporal adverbial modifier for an action

verb.

Example 6(a): *CD4+ T cells were isolated from ST samples, as previously described [27]*. **Briefly,** fresh ST samples were fragmented and digested with collagenase and DNase for 1 hour at 37°C. (Restatement.Specification)

Example 6(b): 2.5×10^6 cells were lysed in lysis buffer [100 mM N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid (HEPES), pH 7.9, 10 mM KCl, 0.1 mM EDTA, 1.5 mM MgCl2, 0.2% Nonidet P-40, 1 mM dithiothreitol (DTT), and 0.5 mM PMSF], **briefly** vortexed at a moderate speed, then incubated on ice for 5 minutes.

2. Materials

The following table below shows the connectives and their frequency as connectives in the BioDRB corpus.

Connective	**Frequency**	**% as Connectives**
and	273	8.17%
by	259	26.26%
To	209	10.80%
after	150	64.11%
however	117	100.00%
also	93	46.97%
then	93	93.00%
thus	77	95.10%
although	77	100.00%
therefore	75	98.68%
when	65	73.86%
while	60	93.75%
whereas	59	98.33%
since	52	73.86%
in contrast	48	93.75%
as	44	5.31%
But	43	18.53%
following	41	56.16%
furthermore	41	100.00%
because	36	83.72%
In order to	36	100.00%
due to	30	63.83%
through	24	34.78%
before	24	61.53%
moreover	22	100.00%
if	21	36.84%
Finally	21	87.5%
On	17	3.68%
via	15	7.65%
followed by	15	48.38%
upon	15	60.00%

indeed	15	83.33%
for example	14	82.35%
prior to	14	100.00%
subsequently	13	86.67%
Briefly	12	60.00%
in response to	11	26.83%
until	11	68.75%
or	10	0.2%
i.e.	10	52.63%
During	9	12.67%
next	9	60.00%
additionally	9	100.00%
in fact	8	72.73%
On the contrary	8	100.00%
alternatively	8	100.00%
once	7	5.26%
further	7	6.93%
except	7	43.75%
so	7	46.67%
thereby	7	100.00%
in turn	6	75.00%
In brief	6	100.00%
Consequently	6	100.00%
on the other hand	6	100.00%
still	5	23.81%
similarly	5	27.77%
albeit	5	83.33%
nevertheless	5	83.33%
as a result of	5	83.33%
and/or	4	12.13%
In addition to	4	30.77%
as a consequence	4	57.14%
in particular	4	57.14%
by the fact that	4	80.00%
such that	4	80.00%
by contrast	4	100.00%
for	3	0.2%
Second	3	6.00%
later	3	23.10%
not only	3	30.00%
whilst	3	60.00%
unless	3	60.00%
in part by	3	75.00%
namely	3	75.00%
In summary	3	75.00%
As an example	3	100.00%
nonetheless	3	100.00%
with	2	0.15%
Given	2	10.00%

rather	2	11.76%
On the basis of	2	22.22%
in that	2	22.22%
but also	2	22.22%
thereafter	2	33.33%
Regardless of	2	66.67%
by means of	2	66.67%
accordingly	2	66.67%
probably	2	100.00%
both upon	2	100.00%
in large part	2	100.00%
for instance	2	100.00%
Instead	2	100.00%
either	1	1.33%
Based on	1	2.86%
In comparison with	1	3.33%
hereafter	1	9.10%
Notably	1	12.50%
though	1	14.29%
e.g.	1	16.67%
mainly by	1	16.67%
nor	1	16.67%
besides	1	20.00%
Third	1	25.00%
as demonstrated by	1	25.00%
in conclusion	1	33.33%
In general	1	50.00%
primarily by	1	50.00%
In view of the fact that	1	100.00%
In outline	1	100.00%
Provided that	1	100.00%
In an effort to	1	100.00%
In order for	1	100.00%
Despite	1	100.00%
appears to be at least		100.00%
in part to	1	
in part via	1	100.00%
Conversely	1	100.00%
insofar as	1	100.00%
as inferred by	1	100.00%
in short	1	100.00%
Now that	1	100.00%
meanwhile	1	100.00%
specifically to	1	100.00%

Table 1: Connectives and their frequencies in BioDRB

3. Learning Features

The following syntactic features were used:

- **Part-of-speech tag (POS)**: The part-of-speech tag associated with the token. For example, in (PP (IN In) (NP (NN contrast))), IN and NN are used as the POS features for the word tokens "in" and "contrast."
- **Parent Category**: The category of the immediate parent of the POS of the token. In the above example, the value of the parent category feature for the token “in” is PP and “contrast” is NP.
- **Left Sibling**: The POS of the token to the left of the current word inside the innermost constituent. If the left POS does not exist (i.e., if the token is the first token in the parent category), the feature will have NONE as its value. In the example above, the value of the left sibling feature for both “in” and “contrast” is NONE, assuming they are the start of a new sentence.

The domain-specific features explored are:

- **Unified Medical Language System (UMLS) Semantic type**: Metamap (http://metamap.nlm.nih.gov/), a program that extracts UMLS concepts associated with the text was applied to obtain the UMLS semantic types. All the semantic types associated with the token were added as a feature.
- **Gene Category**: The BANNER gene tagger was applied to obtain all the mentions of gene in the text and added as a feature.
- **Species Category**: The LINNAEUS species tagger was applied. All the instances tagged as species names were added as a feature.

4. Results of experiments in Open-domain

Table 2 shows the precision, recall, and F1 score of the classifier trained on the 0.24, 0.48, 0.7, and 1 million tokens of the PDTB corpus. Automatically generated syntactic features reduce the performance as compared to human annotated syntactic features. As shown in Table 2, the performance of 0.24 million token data set, the F1 score decreases from 0.918 to 0.829, and this difference is statistically significant ($p<0.01$, t-test, two tails).

	Stanford parser - 0.24 million tokens	**Gold syntax - 0.24 million tokens**	**Gold syntax - 0.48 million tokens**	**Gold syntax - 0.7 million tokens**	**Gold syntax - 1 million tokens**
Precision	0.875 ± 0.018	0.935 ± 0.021	0.938 ± 0.012	0.944 ± 0.007	0.935 ± 0.016
Recall	0.789 ± 0.034	0.902 ± 0.026	0.923 ± 0.010	0.931 ± 0.008	0.925 ± 0.017
F1 score	0.829 ± 0.021	0.918 ± 0.015	0.931 ± 0.008	0.937 ± 0.004	0.930 ± 0.012

Table 2: The performance (average±Std) of *Open-domain* classifier for identifying discourse connectives on different data sizes.

Experiments were performed to ascertain the value of syntactic features with the Open-domain classifier. Starting with just the default ABNER feature set, each syntactic feature was considered independently and then various combinations of features were evaluated. Table 3 shows the precision, recall, and F1 score of the ten-fold cross-validation results of this experiment. The Parent Category feature is the single most effective feature, resulting in an F1 score of 0.922, while the Left Sibling feature also improves performance. Alone, the POS tag feature has a very small effect on the performance. In experiments with combined syntactic features, the POS tag feature still seems to be the

least effective. The performance of all features combined is the same as the performance of all features, except POS tag. The Left Sibling feature improves performance in combination with the Parent category feature (from 0.922 F1 score to 0.930), but the Parent category feature appears from these experiments to be the most valuable by far.

	Precision	**Recall**	**F1 score**
Default	0.876±0.021	0.809±0.012	0.841±0.013
Default+POS	0.878±0.018	0.810±0.024	0.842±0.016
Default+Parent	0.932±0.016	0.914±0.018	0.922±0.013
Default+LeftSib	0.908±0.020	0.875±0.018	0.891±0.012
Default+POS+LeftSib	0.897±0.032	0.875±0.020	0.890±0.014
Default+POS+Parent	0.934±0.016	0.917±0.020	0.926±0.013
Defaul+LeftSib+Parent	0.936±0.017	0.925±0.019	0.930±0.012
All Features	0.935±0.016	0.925±0.017	0.930±0.012

Table 3: Performance (average±Std) of *Open-domain* classifier with combinations of syntactic features

Default+POS: Default features of ABNER + POS feature

Default+Parent: Default features of ABNER + Parent Category feature

Default+LeftSib: Default features of ABNER + Left Sibling feature

Default+POS+LeftSib: Default features of ABNER + POS + Left Sibling

Default+POS+Parent: Default features of ABNER + POS + Parent Category

Default+LeftSib+Parent: Default features of ABNER + Left Sibling + Parent Category

Syntactic feature selection experiments (Table 3) show that the parent category feature had the highest single impact on the performance of the classifier compared with the other two syntactic features. The POS category had the least impact on the performance of the classifier, suggesting that POS information is largely redundant with the

information about the word itself and not very useful. On the other hand, POS features may be valuable for new target domains, as they may help identify previously unseen connectives. The un-adapted source-domain data may thus hurt adaptation performance by reducing the weight of POS features. Future work will continue to explore POS features and related syntactic features and their benefit to the biomedical domain.

5. Error Analysis

To ascertain the effect of singleton connectives on the performance of the classifier, all singleton connectives were removed and an *In-domain* classifier was built for this set. The classifier obtained an F1-score of 0.763 for twelve-fold cross-validation. The difference in the performance of this classifier and the *In-domain* classifier with singleton connectives was not statistically significant ($p<0.01$, t-test, two tails).

The connectives *by* and *to* are highly ambiguous and are not annotated in the PDTB corpus. The connective *by* appears as Noun Phrase (NP) sometimes and as Clause introduced as a subordinating conjunction (SBAR) few times. In either case it may or may not be a connective; therefore, the connectives *by* and *to* were removed in the modified gold standard. Since the removal of singleton connectives did not have significant effect on the result, they were also removed in the modified set of data.

Experiments were then performed on this modified set of data and classifiers that satisfy the two criteria described earlier were built. The results of the experiments are shown in Table 4. The overall performance of all the classifiers increased significantly except for *Weighted-FeatAugment*.

The performance of the *Cross-Domain* classifier increased significantly to 0.673. This

increase is due to the removal of connectives *by* and *to*, which are highly ambiguous and not annotated in the PDTB corpus. The *Unweighted* classifier had a performance of 0.766.

The *In-domain* classifier had an F1-score of 0.791. The performance of all classifiers using simple domain adaptation techniques increased with the *FeatAugment* classifier performing as well as *In-domain* with an F1-score of 0.791. *Weighted* and *Pruning* classifiers had an F1 score of 0.770 and 0.718, respectively.

The performance of the combined domain adaptation techniques also improved except for *Weighted-FeatAugment*, which performed only as well as the *Cross-domain*. The performance of *Weighted-Pruning*, *Weighted-FeatAugment*, *Hybrid*, and *Weighted-Hybrid* are 0.788, 0.690, 0.792, and 0.789 respectively. The *Hybrid* classifier still had the best performance.

Classifier Type	Precision	Recall	F1 Score
Cross-domain	0.824 ± 0.057	0.570 ± 0.064	0.673 ± 0.058
UnWeighted	0.826 ± 0.063	0.715 ± 0.068	0.766 ± 0.059
In-domain	0.846 ± 0.060	0.746 ± 0.074	0.791 ± 0.056
Weighted	0.825 ± 0.068	0.725 ± 0.062	0.770 ± 0.053
Pruning	**0.847 ± 0.060**	0.625 ± 0.072	0.718 ± 0.062
FeatAugment	0.835 ± 0.056	0.755 ± 0.072	0.791 ± 0.054
Weighted-Pruning	0.835 ± 0.061	0.750 ± 0.079	0.788 ± 0.060
Weighted-FeatAugment	0.824 ± 0.067	0.596 ± 0.073	0.690 ± 0.068
Hybrid	0.836 ± 0.058	**0.757 ± 0.074**	**0.792 ± 0.053**
Weighted-Hybrid	0.839 ± 0.055	0.749 ± 0.073	0.789 ± 0.053

Table 4: Performance (average±Std) of different classifiers for identifying the discourse connectives without singleton connectives and connectives *by* and *to*.