Towards Cross-Domain PDTB-Style Discourse Parsing

Evgeny A. Stepanov and **Giuseppe Riccardi**
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, Italy
{stepanov,riccardi}@disi.unitn.it

Abstract

Discourse relation parsing is an important task with the goal of understanding text beyond the sentence boundaries. With the availability of annotated corpora (Penn Discourse Treebank) statistical discourse parsers were developed. In the literature it was shown that the discourse parsing subtasks of discourse connective detection and relation sense classification do not generalize well across domains. The biomedical domain is of particular interest due to the availability of Biomedical Discourse Relation Bank (BioDRB). In this paper we present cross-domain evaluation of PDTB trained discourse relation parser and evaluate feature-level domain adaptation techniques on the argument span extraction subtask. We demonstrate that the subtask generalizes well across domains.

1 Introduction

Discourse analysis is one of the most challenging tasks in Natural Language Processing that has applications in many language technology areas such as opinion mining, summarization, information extraction, etc. (see (Webber et al., 2011) and (Taboada and Mann, 2006) for detailed review). The release of the large discourse relation annotated corpora, such as Penn Discourse Treebank (PDTB) (Prasad et al., 2008), marked the development of statistical discourse parsers (Lin et al., 2012; Ghosh et al., 2011; Xu et al., 2012; Stepanov and Riccardi, 2013). Recently, PDTB-style discourse annotation was applied to biomedical domain and Biomedical Discourse Relation Bank (BioDRB) (Prasad et al., 2011) was released. This milestone marks the beginning of the research on cross-domain evaluation and domain adaptation of PDTB-style discourse parsers. In this paper we address the question of how well PDTB-trained discourse parser (news-wire domain) can extract argument spans of *explicit* discourse relations in BioDRB (biomedical domain).

The use cases of discourse parsing in biomedical domain are discussed in detail in (Prasad et al., 2011). Here, on the other hand, we provide very general connection between the two. The goal of Biomedical Text Mining (BioNLP) is to retrieve and organize biomedical knowledge from scientific publications; and detecting discourse relations such as contrast and causality is an important step towards this goal (Prasad et al., 2011). To illustrate this point consider a quote from (Brunner and Wirth, 2006), given below.

> *The addition of an anti-Oct2 antibody did not interfere with complex formation (Figure 3, lane 6)*, since **HeLa cells do not express Oct2**. (Cause:Reason)

In the example, the discourse connective since signals a causal relation between the clauses it connects. That is, the reason why '*the addition of an anti-Oct2 antibody did not interfere with complex formation*' is '*HeLa cells' not expressing Oct2*'.

PDTB adopts non-hierarchical binary view on discourse relations: Argument 1 (*Arg1*) (in italics in the example) and Argument 2 (*Arg2*), which is syntactically attached to a discourse connective (in bold). Thus, a discourse relation is a triplet of a connective and its two arguments. In the literature (Lin et al., 2012; Stepanov and Riccardi, 2013) PDTB-style discourse parsing is partitioned into discourse relation detection, argument position classification, argument span extraction, and relation sense classification. For the *explicit* discourse relations (i.e. signaled by a connective), discourse relation detection is cast as classification of connectives as discourse and non-discourse. Argument position classification, on the other hand, involves detection of the location of *Arg1* with re-

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 30–37,
Gothenburg, Sweden, April 26-30 2014.

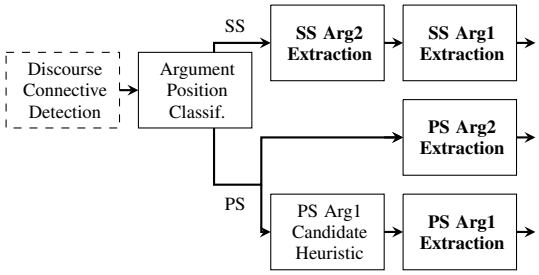

Figure 1: Discourse Parser Architecture. (CRF Argument Span Extraction models are in bold.)

spect to *Arg2*, that is to detect whether a relation is inter- or intra- sentential. Argument span extraction is the extraction (labeling) of text segments that belong to each of the arguments. Finally, relation sense classification is the annotation of relations with the senses from the sense hierarchy (PDTB or BioDRB).

To the best of our knowledge, the only subtasks that were addressed cross-domain are the detection of explicit discourse connectives (Ramesh and Yu, 2010; Ramesh et al., 2012; Faiz and Mercer, 2013) and relation sense classification (Prasad et al., 2011). While the discourse parser of Faiz and Mercer (2013)[1] provides models for both domains and does identification of argument head words in the style of Wellner and Pustejovsky (2007); there is no decision made on arguments spans. Moreover, there is no cross-domain evaluation available for each of the models. In this paper we address the task of cross-domain argument span extraction of *explicit* discourse relations. Additionally, we provide evaluation for cross-domain argument position classification as far as the data allows, since BioDRB lacks manual sentence segmentation.

The paper is structured as follows. In Section 2 we present the comparative analysis of PDTB and BioDRB corpora and the relevant works on cross-domain discourse parsing. In Section 3 we describe the PDTB discourse parser used for cross-domain experiments. In Section 4 we present the evaluation methodology and the experimental results. Section 5 provides concluding remarks.

2 PDTB vs. BioDRB Corpora Analysis and Related Cross-Domain Works

The two corpora used in our experiments are Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and Biomedical Discourse Relation Bank (BioDRB) (Prasad et al., 2011). Both corpora follow the same discourse relation annotation style over different domain corpora: PDTB is annotated on top of *Wall Street Journal* (WSJ) corpus (financial news-wire domain); and it is aligned with Penn Treebank (PTB) syntactic tree annotation; BioDRB, on the other hand, is a corpus annotated over 24 open access full-text articles from the GENIA corpus (Kim et al., 2003) (biomedical domain), and, unlike PDTB, there is no reference tokenization or syntactic parse trees.

The detailed comparison of the corpora is out of the scope of this paper, and it is available in (Prasad et al., 2011). Similarly, the review of PDTB-style discourse parsing literature is not in its scope. Here, on the other hand, we focus on the corpus differences relevant for discourse parsing tasks and cross-domain application of discourse parsing subtasks.

Discourse relations in both corpora are binary: *Arg1* and *Arg2*, where *Arg2* is an argument syntactically attached to a discourse connective. With respect to *Arg2*, *Arg1* can appear in the same sentence (SS case), one or several of the preceding (PS case) or following (FS case) sentences. A discourse connective is a member of a well defined list of connectives and a relation expressed via such connective is an *Explicit* relation. There are other types of discourse and non-discourse relations annotated in the corpora; however, they are out of the scope of this paper. Discourse relations are annotated using a hierarchy of senses: even though the organization of senses and the number of levels are different between corpora, the most general top level senses are mapped to the PDTB top level senses: *Comparison*, *Contingency*, *Expansion*, and *Temporal* (Prasad et al., 2011).

The difference between the two corpora with respect to discourse connectives is that in case of PDTB the annotated connectives belong to one of the three syntactic classes: subordinating conjunctions (e.g. *because*), coordinating conjunctions (e.g. *but*), and discourse adverbials (e.g. *however*), while BioDRB is also annotated for a forth syntactic class – subordinators (e.g. *by*).

There are 100 unique connective types in PDTB (after connectives like *1 year after* are stemmed to *after*) in 18,459 explicit discourse relations. Whereas in BioDRB there are 123 unique connective types in 2,636 relations. According to

[1]Made available on `https://code.google.com/p/discourse-parser/`

the discourse connective analysis in (Ramesh et al., 2012), the subordinators comprise 33% of all connective types in BioDRB. Additionally, 11% of connective types in common syntactic classes that occur in BioDRB do not occur in PDTB; e.g. *In summary*, *as a consequence*. Thus, only 56% of connective types of BioDRB are common to both corpora. While in-domain discourse connective detection has good performance (Ramesh and Yu, 2010), this difference makes the cross-domain identification of discourse connectives a hard task, which is exemplified by experiments in (Ramesh and Yu, 2010) ($F_1 = 0.55$).

With respect to relation sense classification, the connective surface provides already high baselines (Prasad et al., 2011). However, cross-domain sense classification experiments indicate that there are significant differences in the semantic usage of connectives between two domains, since the performance of the classifier trained on PDTB does not generalize well to BioDRB ($F_1 = 0.57$).

To sum up, the corpora differences with respect to discourse connective usage affect the cross-domain generalization of connective detection and sense classification tasks negatively. The experiments in this paper are intended to evaluate the generalization of argument span extraction, assuming that the connective is already identified. In the following section, we present the PDTB-trained discourse parser optimized for in-domain performance.

3 PDTB-Style Discourse Parser

The discourse parser (see Figure 1) is a combination of argument position classification model for classifying discourse connectives as inter- or intra-sentential, and specific Conditional Random Fields argument extraction models for each of the arguments in these configurations. In the following subsections we provide descriptions for each of the components.

3.1 Argument Position Classification

Discourse connectives have a very strong preference on the location of the *Arg1* with respect to their syntactic category (Subordinating Conjunction, Coordinating Conjunction, and Discourse Adverbial) and position in the sentence (sentence initial or sentence medial); thus, classification of discourse connectives into inter-sentential or intra-sentential is an easy task yielding high supervised machine learning performance (Stepanov and Riccardi, 2013; Lin et al., 2012). With respect to the decision made in this step a specific argument span extraction model is applied.

For *Argument Position Classification* the unigram BoosTexter (Schapire and Singer, 2000) model with 100 iterations is trained on PDTB sections 02-22 and tested on sections 23-24. Similar to the previously published results, it has a high performace: $F1 = 98.12$. The features are connective surface string, POS-tags, and IOB-chains. The results obtained with automatic sentence splitting, tokenization, and syntactic parsing using Stanford Parser (Klein and Manning, 2003) are also high $F1 = 97.81$.

Since, unlike PTB for PDTB, for BioDRB there is no manual sentence splitting, tokenization, and syntactic tree annotation; the precise cross-domain evaluation of *Argument Span Extraction* step is not possible. However, in Section 4 we estimate the performance using automatic sentence splitting.

3.2 Argument Span Extraction

Argument span extraction is cast as token-level sequence labeling using Conditional Random Fields (CRF) (Lafferty et al., 2001). Previously, it was observed that in PDTB for inter-sentential discourse relations *Arg1* precedes *Arg2* in most of the cases. Thus, the CRF models are trained for the configurations where both of the arguments are in the same sentence (SS), and for *Arg1* in one of the previous sentences (PS); the following sentence *Arg1* case (FS) is ignored due to too few training instances being available (in PDTB 8 / 18,459). Consequently, there are 4 CRF models SS Arg1 and Arg2, and PS Arg1 and Arg2.

Same sentence case models are applied in a cascade, such that output of *Arg2* model is used as a feature for *Arg1* span extraction. For the case of *Arg1* in the previous sentences; based on the observation that in PDTB Arg2 span is fully located in the sentence containing the connective in 98.5% of instances; and Arg1 span is fully located in the sentence immediately preceding Arg2 in 71.7% of instances; the sentences in these positions are selected and CRF models are trained to label the spans.

The features used for training the models are presented in Table 1. The feature sets are optimized for each of the arguments in (Ghosh et al., 2011) (see the Table columns Arg1 and Arg2). Be-

sides the features commonly used in NLP tasks such that token, lemma, inflectional affixes, and part-of-speech tag, the rest of the features are:

- *IOB-Chain* (*IOB*) is the path string of the syntactic tree nodes from the root node to the token, prefixed with the information whether a token is at the beginning (B-) or inside (I-) the constituent. The *chunklink* tool (Buchholz, 2000) is used to extract this feature from syntactic trees.
- *PDTB Level 1 Connective sense* (*CONN*) is the most general sense of a connective in PDTB sense hierarchy. It's general purpose is to label the discourse connective tokens, i.e. the value of the feature is '*NULL*' for all tokens except the discourse connective.
- *Boolean Main Verb* (*BMV*) is a boolean feature that indicates whether a token is a main verb of a sentence or not (Yamada and Matsumoto, 2003).
- *Arg2 Label* (*ARG2*) is an output of *Arg2* span extraction model, that is used as a feature for *Arg1* span extraction. *Arg2* span is easier to identify (Ghosh et al., 2011; Stepanov and Riccardi, 2013) since it is syntactically attached to the discourse connective. Thus, this feature serves to constrain the *Arg1* search space for intra-sentential argument span extraction. The value of the feature is either ARG2 suffixed for whether a token is Inside (I), Begin (B), or End (E) of the span, or 'O' if it does not belong to the *Arg2* span.

These features are expanded during training with n-grams (feature of CRF++[2]): tokens with 2-grams in the window of ± 1 tokens, and the rest of the features with 2 & 3-grams in the window of ± 2 tokens.

The in-domain performance of argument span extraction models is provided in the following section, after the description of the evaluation methodology.

4 Experiments and Results

In this Section we first describe the evaluation methodology and then the experiments on cross-domain evaluation of argument position classification and argument span extraction models.

[2]https://code.google.com/p/crfpp/

Feature	ABBR	Arg2	Arg1
Token	TOK	Y	Y
POS-Tag	POS		
Lemma	LEM	Y	Y
Inflection	INFL	Y	Y
IOB-Chain	IOB	Y	Y
Connective Sense	CONN	Y	Y
Boolean Main Verb	BMV		Y
Arg2 Label	ARG2		Y

Table 1: Feature sets for Arg2 and Arg1 argument span extraction.

The experimental settings for PDTB are the following: Sections 02-22 are used for training and Sections 23-24 for testing. For BioDRB, on the other hand, 12 fold cross-validation is used (2 documents in each fold, since in BioDRB there are 24 documents).

4.1 Evaluation Methodology

The performance of *Argument Span Extraction* is evaluated in terms of precision (*p*), recall (*r*), and F-measure (F_1) using the equations 1 – 3. An argument span is considered to be correct, if it exactly matches the reference string. Following (Ghosh et al., 2011) and (Lin et al., 2012), argument initial and final punctuation marks are removed .

$$p = \frac{\text{Exact Match}}{\text{Exact Match} + \text{No Match}} \quad (1)$$

$$r = \frac{\text{Exact Match}}{\text{References in Gold}} \quad (2)$$

$$F_1 = \frac{2 * p * r}{p + r} \quad (3)$$

In the equations, *Exact Match* is the count of correctly tagged argument spans; *No Match* is the count of argument spans that do not match the reference string exactly, i.e. even a single token difference is counted as an error; and *References in Gold* is the total number of arguments in the reference.

Since argument span extraction is applied after argument position classification, the classification error is propagated. Thus, for the evaluation of argument span extraction, misclassified instances are reflected in the counts of *Exact Matches* and *No Matches*. For example, misclassified same sentence relation results in that both its arguments are

	Arg2			Arg1		
	P	**R**	**F1**	**P**	**R**	**F1**
Gold						
SS	90.36	87.49	88.90	70.27	66.67	68.42
PS	79.01	77.10	78.04	46.23	36.61	40.86
ALL	85.93	83.45	84.67	61.94	54.98	58.25
Auto						
SS	86.83	85.14	85.98	64.26	63.01	63.63
PS	75.00	73.67	74.33	37.66	37.00	37.33
ALL	82.24	80.69	81.46	53.93	52.92	53.42

Table 2: In-domain performance of the PDTB-trained argument span extraction models on the test set with 'Gold' and 'Automatic' sentence splitting, tokenization, and syntactic features. The results are reported together with the error propagation from argument position classification for Same Sentence (SS), Previous Sentence (PS) models and joined results (ALL) as precision (P), recall (R) and F-measure (F1).

considered as not recalled for the SS, and for the PS they are considered as *No Match*.

However, we do not propagate error in cross-domain evaluation on BioDRB, since there is no reference information. Additionally, while *Arg1* span extraction models are trained on Gold *Arg2* features, for testing they are always automatic.

4.2 Cross-Domain Argument Position Classification

As it was mentioned above, there is no manual sentence splitting for BioDRB; thus, there is no references for whether a discourse relation has its *Arg1* in the same or different sentences. In order to evaluate cross-domain argument position classification we evaluate classifier decisions against automatic sentence splitting using Stanford Parser (Klein and Manning, 2003) on whole of BioDRB.

The BoosTexter model described in Section 3.1 has a high in-domain performance of 97.81. On BioDRB its performance is 95.26, which is still high. Thus, we can conclude that argument position classification generalizes well cross-domain, and that it is little affected by the presence of 'subordinators' that were not annotated in PDTB.

4.3 In-Domain Argument Span Extraction: PDTB

The in-domain performance of the argument span extraction models trained on PDTB sections 02-22 and tested on sections 23-24 is given on Table 2. The results are for 2 settings: 'Gold' and 'Auto'. In the 'Gold' settings the sentence splitting, tokenization and syntactic features are extracted from PTB, and in the 'Auto' they are extracted from automatic parse trees obtained using Stanford Parser (Klein and Manning, 2003).

The general trend in the literature, is that the argument span extraction for *Arg1* has lower performance than for *Arg2*, which is expected since *Arg2* position is signaled by a discourse connective. Additionally, Previous Sentence *Arg1* model performance is much lower than that of the other models due to the fact that it only considers immediately previous sentence; which, as was mentioned earlier, covers only 71.7% of the inter-sentential relations. In the next subsections, these models are evaluated on biomedical domain.

4.4 In-Domain Argument Span Extraction: BioDRB

In order to evaluate PDTB-BioDRB cross-domain performance we first evaluate the in-domain BioDRB argument span extraction. Since there is no gold sentence splitting, tokenization and syntactic parse trees, the models are trained using the features extracted from automatic parse trees. We use exactly the same feature sets as for PDTB models, which are optimized for PDTB. An important aspect is that in BioDRB the connective senses are different: there are 16 top level senses that are mapped to 4 top level PDTB senses. For the in-domain BioDRB models, the 16 senses were kept as is.

Since we do not have gold argument position information, we do not train in-domain argument classification model. Thus, the reported results are without error propagation. Later, this will allow us to assess cross-domain argument span extraction performance better.

The results reported in Table 3 are average precision, recall and f-measure of 12-fold cross-validation. With respect to automatic sentence splitting, there are 717 inter-sentential and 1,919 intra-sentential relations (27% to 73%). Thus, BioDRB is less affected by PS *Arg1* performance than PDTB models, where the ratio is 619 to 976 (39% to 61%). Additionally, BioDRB PS *Arg1* performance is generally higher than that of PDTB. Overall, in-domain BioDRB argument extraction model performance is in-line with the

	Arg2			Arg1		
	P	**R**	**F1**	**P**	**R**	**F1**
SS	80.94	79.88	80.41	66.51	61.82	64.07
PS	82.99	82.99	82.99	57.50	55.62	56.53
ALL	81.45	80.67	81.06	63.87	60.00	61.87

Table 3: In-domain performance of the BioDRB-trained argument span extraction models. Both training and testing are on automatic sentence splitting, tokenization, and syntactic features. The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation.

PDTB models, with the exception that previous sentence *Arg2* has higher performance than the same sentence one.

4.5 Cross-Domain Argument Span Extraction: PDTB - BioDRB

Similar to in-domain BioDRB argument span extraction, we perform 12 fold cross-validation for PDTB-BioDRB cross-domain argument span extraction. The cross-domain performance of the models described in Section 4.3 is given in the Table 4 under the 'Gold'. To make the cross-domain evaluation settings closer to the BioDRB in-domain evaluation, we additionally train PDTB models on the automatic features, i.e. features extracted from PDTB with automatic sentence splitting, tokenization and syntactic parsing. Similar to the in-domain BioDRB evaluation, results are reported without error propagation from argument position classification step.

The first observation from cross-domain evaluation is that argument span extraction generalizes to biomedical domain much better that the discourse parsing subtasks of discourse connective detection and relation sense classification. Unlike those subtasks, the difference between in-domain BioDRB argument span extraction models and the models trained on PDTB is much less: e.g. for discourse connective detection the in-domain and cross-domain difference for BioDRB is 14 points (f-measures 69 and 55 in (Ramesh and Yu, 2010)), and for argument span extraction 2 and 4 points for *Arg2* and *Arg1* respectively (see Tables 3 & 4).

The difference between the models trained on automatic and gold parse trees is also not high, and gold feature trained models perform better with

	Arg2			Arg1		
	P	**R**	**F1**	**P**	**R**	**F1**
Gold						
SS	80.37	76.58	78.42	60.82	56.40	58.52
PS	80.73	80.50	80.62	57.74	52.95	55.19
ALL	80.53	77.71	79.09	59.76	55.29	57.43
Auto						
SS	77.60	75.05	76.30	60.76	55.21	57.83
PS	81.39	81.23	81.31	57.71	51.72	54.47
ALL	78.72	76.80	77.74	59.60	54.12	56.71

Table 4: Cross-domain performance of the PDTB-trained argument span extraction models on BioDRB. For the 'Gold' setting the models from in-domain PDTB section are used. For 'Auto', the models are trained on automatic sentence splitting, tokenization, and syntactic features. The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation.

the exception of PS *Arg2*. Since training on automatic parse trees does not improve cross-domain performance, the rest of the experiments is using gold features for training.

4.6 Feature-Level Domain Adaptation

The two major differences between PDTB and BioDRB are vocabulary and connective senses. The out-of-vocabulary rate of PDTB on the whole BioDRB is 22.7% and of BioDRB on PDTB is 33.1%, which are very high. Thus, PDTB lexical features might not be very effective, and the models generalize well due to syntactic features. To test this hypothesis we train additional PDTB models on only syntactic features: POS-tags and IOB-chain and 'connective labels' – 'CONN' suffixed for the Beginning (B), Inside (I) or End (E) of the connective span, simulating discourse connective detection output. Moreover, we reduce the feature set to **unigrams** only (recall that features were enriched by 2 and 3 grams), such that the models become very general.

Even though BioDRB connective senses can be mapped to PDTB, in (Prasad et al., 2011) it was observed that relation sense classification does not generalize well. To reduce the dependency of argument span extraction models on relation sense classification, the connective sense feature in the

	Arg2			Arg1		
	P	**R**	**F1**	**P**	**R**	**F1**
Baseline						
SS	80.37	76.58	78.42	60.82	56.40	58.52
PS	80.73	80.50	80.62	57.74	52.95	55.19
ALL	80.53	77.71	79.09	59.76	55.29	57.43
Syntactic						
SS	82.00	75.03	78.33	61.07	51.80	56.01
PS	75.56	74.47	75.01	56.64	46.66	51.11
ALL	80.31	74.98	77.54	59.69	50.42	54.63
No Relation Sense						
SS	81.35	74.00	77.47	62.46	56.11	59.10
PS	80.35	80.13	80.24	57.58	52.25	54.74
ALL	81.16	75.67	78.30	60.86	54.87	57.69

Table 5: Cross-domain performance of the PDTB-trained argument span extraction models on BioDRB. For the 'Syntactic' setting the models are trained on only syntactic features (POS-tag + IOB-chain) and 'connective labels'. For 'No Relation Sense', the models are trained by replacing connective sense with 'connective labels'. The 'Baseline' is repeated from Table 4. The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation.

'Baseline' models (i.e. the models from Section 4.3) is also replaced by 'connective labels'. We train these models using gold features only, and, similar to previous experiments, do 12-fold cross-validation.

The performance of the adapted models is given in Table 5. The 'Syntactic' section gives the results of the models trained on syntactic features and the 'No Relation Sense' section gives the results for the models with 'connective labels' instead of connective senses, and the 'Baseline' repeats the performance of the PDTB-optimized models.

The PDTB-optimized baseline, outperforms the adapted models on *Arg2*; however, 'No Relation Sense' *Arg1* yields the best performance, and, though insignificantly, outperforms the baseline. Thus, the effect of replacing connective senses with 'connective labels' is negative for all cases except SS *Arg1*. Overall, the difference in performance between the 'Baseline' and 'No Relation Sense' models is an acceptable price to pay for the independence from relation sense classification.

	Arg2			Arg1		
	P	**R**	**F1**	**P**	**R**	**F1**
SS	81.72	76.14	78.82	61.53	56.36	58.82
PS	80.31	79.84	80.07	58.55	52.82	55.44
ALL	81.27	77.10	79.12	60.56	55.30	57.80

Table 6: Cross-domain performance of the PDTB-trained argument span extraction model on unigram and bigrams of token, POS-tag, IOB-chain and 'connective label'. The results are reported for Same Sentence (SS) and Previous Sentence (PS) models, and the joined results for each of the arguments (ALL) as average precision (P), recall (R), and F-measure (F1) of 12-fold cross-validation.

The most general models – unigrams of Part-of-Speech tags and IOB-chains together with 'connective labels' in the window of ± 2 tokens – all have the performance lower than the baseline, which is expected given its feature set. However, for the easiest case of intra-sentential *Arg2* it outperforms the model trained by replacing the connective sense in the baseline (i.e. 'No Relation Sense'). Degraded performance of *Arg1* models indicates that lexical features are helpful.

Introducing the tokens back into the 'Syntactic' model, and increasing the features to include also 2-grams, boosts the performance of the models to outperform the 'No Relation Sense' models in all but Previous Sentence *Arg2* category. However, the models now yield performance comparable to the PDTB optimized baseline (insignificantly better), while being unaffected by poor cross-domain generalization of relation sense classification (see Table 6).

The cross-domain argument extraction experiments indicate that models trained on PDTB-optimized feature set already have good generalization. However, they are dependent on relation sense classification task, which does not generalize well. By replacing connective senses with 'connective labels' we obtain models independent of this task while maintaining comparable performance. The in-domain trained BioDRB models, however, perform better, as expected.

5 Conclusion

In this paper we presented cross-domain discourse parser evaluation on subtasks of argument position classification and argument span extraction.

The observed cross-domain performances are indicative of good model generalization. However, since these models are applied later in the pipeline, they are affected by the cross-domain performance of the other tasks. Specifically, discourse connective detection, which was shown not to generalize well in the literature. Additionally, we have presented feature-level domain adaptation techniques to reduce the dependence of the cross-domain argument span extraction on other discourse parsing subtasks.

The syntactic parser (Stanford) that provides sentence splitting and tokenization is trained on Penn Treebank, i.e. it is in-domain for PDTB and out-of-domain for BioDRB; and it is known that domain-optimized tokenization improves performance on various NLP tasks. Thus, the future direction of this work is to evaluate argument span extraction using tools optimized for biomedical domain.

Acknowledgments

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 610916 – SENSEI.

References

Cornelia Brunner and Thomas Wirth. 2006. Btk expression is controlled by oct and bob. 1/obf. 1. *Nucleic acids research*, 34(6):1807–1815.

Sabine Buchholz. 2000. Readme for perl script chunklink.pl.

Syeed Ibn Faiz and Robert E Mercer. 2013. Identifying explicit discourse connectives in text. In *Advances in Artificial Intelligence*, pages 64–76. Springer.

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 1:1 – 35.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):188.

Balaji Polepalli Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text. *AMIA Annual Symposium Proceedings*, 2010:657.

Balaji Polepalli Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association*, 19(5):800–808.

Robert E. Schapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

Evgeny A. Stepanov and Giuseppe Riccardi. 2013. Comparative evaluation of argument extraction algorithms in discourse relation parsing. In *13th International Conference on Parsing Technologies (IWPT 2013)*, pages 36–44.

Maite Taboada and William C. Mann. 2006. Applications of rhetorical structure theory. *Discourse Studies*, (8):567–88.

Bonnie L. Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, pages 1 – 54.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*.

Fan Xu, Qiao Ming Zhu, and Guo Dong Zhou. 2012. A unified framework for discourse argument identification via shallow semantic parsing. In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012): Posters*.

Hiroyasu Yamada and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of 8th International Workshop on Parsing Technologies*.