The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Linguistic Properties Matter for Implicit Discourse Relation Recognition: Combining Semantic Interaction, Topic Continuity and Attribution

Wenqiang Lei,[1] **Yuanxin Xiang,**[1] **Yuwei Wang,**[3]
Qian Zhong,[4] **Meichun Liu,**[4] **Min-Yen Kan**[1,2]*
[1]National University of Singapore [2]Smart Systems Institute
[3]University of Utah [4]City University of Hong Kong
{wenqiang, yuanxin, knmnyn}@comp.nus.edu.sg,
ywang@cs.utah.edu,{qzhong5-c,meichliu}@cityu.edu.hk

Abstract

Modern solutions for implicit discourse relation recognition largely build universal models to classify all of the different types of discourse relations. In contrast to such learning models, we build our model from first principles, analyzing the linguistic properties of the individual top-level Penn Discourse Treebank (PDTB) styled implicit discourse relations: *Comparison*, *Contingency* and *Expansion*. We find semantic characteristics of each relation type and two cohesion devices – topic continuity and attribution – work together to contribute such linguistic properties. We encode those properties as complex features and feed them into a Naïve Bayes classifier, bettering baselines (including deep neural network ones) to achieve a new state-of-the-art performance level. Over a strong, feature-based baseline, our system outperforms one-versus-other binary classification by 4.83% for *Comparison* relation, 3.94% for *Contingency* and 2.22% for four-way classification.

1 Introduction

Sentences do not stand alone in text; they must be cohesive by employing some rhetorical device, such as topic continuity and discourse relations. The inventories of discourse relations vary with the particular modeling assumptions adopted in each framework, such as in Rhetorical Structure Theory Treebank (RST) (Carlson, Okurowski, and Marcu 2002) and in the Penn Discourse Treebank (PDTB) (Prasad et al. 2007). In this work, we adopt those of the PDTB, and follow its terminologies (*cf* Section 2.1). A key challenge in computational discourse analysis is the automatic recognition of **implicit** (*i.e.*, relations unmarked by explicit discourse markers) discourse relations (Liu and Li 2016).

Existing solutions focus on creating good universal models, largely ignoring the properties of individual relation types. Some neural network approaches treat the task as a simple classification problem where the input is a pair of sentences (arguments). While this is beneficial in terms of modeling, these models do nothing to advance our understanding to the detailed linguistic properties of each type of implicit discourse relation. We argue that such a general approach is insufficient to address the unique properties of the scenario. Our solution to address these problems can be applied to other tasks such as natural language inference (MacCartney 2009) and text similarity (Agirre et al. 2016).

We recognize that each discourse relation type has its own unique properties, having individual **semantic characteristics**. For example, *Comparisons* often use negation to highlight the contradictory part of two arguments. A typical case is when one argument's predicate is a negated expression of the opposing argument. **Topic continuity** is also an integral device for cohesion (Halliday 1976). It works together with a relation's semantic characteristics to license the discourse relation between two arguments. For example, the arguments for a contradiction (a subtype of the *Comparison* discourse relation in the PDTB) must refer to the same topic. Interestingly, our analyses reveals that topic continuity is manifested differently depending on relation's semantic characteristics. Finally, our analyses also reveal that **attribution** – the source of an argument (*cf* Section 2.1) – is another important but often overlooked cohesion device. As a cohesion device, it provides background context for arguments, providing necessary information for discourse relation. Our analyses show that the collaboration between relations' semantic characteristics and the two cohesion devices of topic continuity and attribution are a hallmark of discourse relationships.

Our study contributes towards the understanding of discourse relations with the goal of improving their automatic recognition. We make the following contributions:

- Through corpus study, we uncover typical patterns in the PDTB that demonstrate the cooperation of discourse relations' semantic characteristics and the two cohesion devices of topic continuity and attribution. By encoding those patterns as complex features, we obtain significant improvement over a strong baseline, achieving a new state-of-the-art level of performance.
- We assert the importance of two specific devices that establish cohesion: topic continuity and attribution. These two cohesion devices are what distinguish implicit discourse relation recognition from other tasks such as natural language inference and text similarity, and we believe their integral role in the implicit discourse recognition has

*This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative, and by NVIDIA Corporation for their donation of a Titan X GPU.

been overlooked by the community.

- In contrast to recent neural models, our model's classification is more transparent, explainable and easy to replicate. We release our source code to aid community adoption of these linguistic insights.[1]

2 Related Work

We briefly review the terminology used by the Penn Discourse Treebank. We then review of existing models on PDTB-styled implicit discourse relation recognition, which have mostly focused on building unified models for all four relations.

2.1 The Penn Discourse Treebank (PDTB)

The PDTB defines four top-level discourse relations: *Comparison*, *Contingency*, *Expansion* and *Temporal*. A key property of PDTB is that it distinguishes between explicit and implicit discourse relations. Explicit relations are overtly signaled by a discourse **connective** (e.g., "but", "however", "because") present in the text; otherwise, the relation is termed implicit.

Table 1 illustrates implicit instances for all four top-level relations. Following the standard PDTB convention, annotations given from PDTB are marked by square brackets with their annotation type indicated by a suffix. **Arguments** are defined as the two text spans between which a discourse relation holds, named **Arg1** and **Arg2**, respectively. For each implicit instance, annotators are asked to infer a suitable connective for each instance which is marked by suffix `conn` in Table 1. As these connectives are inferred, they are only for reference, and not actually present nor provided in training classification methods.

Attributions are also marked by the PDTB. The attribution is the source of arguments. For example, the source span of Arg1 in (Ex. 2.4) is "President Bush insists". Therefore, this span is marked with the suffix `Attr1` (or `Attr2` when the span attributes Arg2). When both Arg1 and Arg2 are attributed to the same source, the source text span is marked `RelAttr`).

2.2 Implicit Discourse Relation Recognition

Existing feature-based work for implicit discourse relation recognition find a group of features to distinguish all four top-level implicit discourse relation. For example, word pairs (Marcu and Echihabi 2002) and Brown cluster pairs (Rutherford and Xue 2014) indicate that the relations' lexical choice influence the inferred discourse relation. General Inquirer tags (Pitler, Louis, and Nenkova 2009) similarly show that the semantic categories of predicates relate to discourse relation types. Parser production rules (Lin, Kan, and Ng 2009) also indicate syntatic patterns also carry the signal of discourse relation. Yet these studies do not explore whether the relation types have specific linguistic properties, with the notable exception of (Rutherford and Xue 2014).

In a separate line of attack on the task, the recent wave of deep learning leverages on neural networks to build universal end-to-end models. For example, Chen *et al.* (2016) develop a Gated Revelance Network (GRN) to capture important word pairs, Liu *et al.* (2016) apply a sophisticated multi-layer attention model, Qin *et al.* (2016) employ recurrent neural networks stacked with convolutional networks, and Lei *et al.* (2017) build a simple word interaction neural model. However, these solutions do not exploit specific properties of each type of discourse relation either, which we argue are necessary to gain linguistic insight on the task of discourse relation identification.

This is not to say that a linguistic basis of analysis has been ignored; rather, it is preliminary and yet to be fully developed. Louis *et al.* (2010) find that cohesion – in the form of coreference features – "do not perform as well as lexical features". Rutherford and Xue (2014) explore the topic continuity of subjects, predicates and objects between arguments, but demonstrate a lackluster 0.6% F_1 score improvement in *Contingency-vs-others* binary classification. Ji and Eisenstein (2015) consider topic continuity when representing arguments into a dense vector, gaining 0.7% F_1 score in four-way classification. We believe these works indicate that cohesion is important for the task, but that existing work has yet to delve into a nuanced analysis which would yield insight and performance gains.

3 Qualitative Corpus Study

Our key finding is that semantic characteristics and two cohesion devices – topic continuity and attribution – coordinate to determine the discourse relation. The specific semantic characteristics vary per discourse relation type, and manifest different patterns of cohesion with respect to topic continuity and attribution. Through our corpus study, we analyze instances of the *Comparison*, *Contingency* and *Expansion* relation types to discover such patterns.

Our study examines certain key representative linguistic phenomena that are important to classification and recognition. We note that there are many other specific linguistic phenomena associated with implicit discourse relations that we do not mention or leverage; it is beyond the scope of this work to build an exhaustive inventory of such phenomena. We now discuss each of the three relation types in turn.

3.1 Comparison

According to the PDTB annotation guidelines (Prasad et al. 2007), a *Comparison* relation "is established between Arg1 and Arg2 in order to highlight prominent differences between the two situations" (*ibid.*). While the "difference" can be expressed in various ways such as antonyms which have been captured by General Inquirer Tags (Pitler, Louis, and Nenkova 2009), negation is a natural device to highlight the difference indicated by contradiction, thus is an important semantic characteristic for the *Comparison* relation.

Negation. Negation for *Comparison* is usually used to express the contradiction towards the same topic. In (Ex. 1.1), the negation introduced by "not" in Arg2 establishes a contradiction towards the topic "prevent" shared by both arguments. This illustrates the importance of the cooperation between negation and topic continuity.

In terms of the semantic characteristic, negation tend to be confined within the predicate of only one of the two

[1] http://github.com/WING-NUS/discoling

Comparison	Ex. (1.1) [To avoid this deficit Mr. Lawson inflated the pound in order to prevent its rise.]$_{Arg1}$ [however]$_{conn}$ [This misguided policy could **not** prevent a British trade deficit.]$_{Arg2}$ WSJ_0571
	Ex. (1.2) [Government lending was **not** intended to be a way to obfuscate spending figures, hide fraudulent activity, or provide large subsidies.]$_{Arg1}$ [instead]$_{conn}$ [The reforms described above would provide a more limited, but clearer, safer and ultimately more useful role for government as a lender.]$_{Arg2}$ WSJ_1131
Contingency	Ex. (2.1) ["Psyllium's not a good crop".]$_{Arg1}$ [**Complains** Sooraji Jath, a 26-year-old farmer from the village of Lakshmipura,]$_{RelaAttr}$ [because]$_{conn}$["You get a rain at the wrong time and the crop is ruined".]$_{Arg2}$ WSJ_0515
	Ex. (2.2) [Carl Schramm, president of the Health Insurance Association of America, **scoffs** at]$_{Attr1}$["capitalists who want to socialize the entire financing system" for health.]$_{Arg1}$[because]$_{conn}$ ["They hope they can buy some government cost discipline," but this is a false hope,]$_{Arg2}$ [Mr. Schramm says.]$_{Attr2}$ WSJ_0314
	Ex. (2.3) [At Applied, Mr. Sim set growth as his first **objective**.]$_{Arg1}$[accordingly]$_{conn}$[He took the company public in an offering that netted Applied about $12.6 million, which helped launch the company's acquisition program.]$_{Arg2}$ WSJ_2282
	Ex. (2.4) [President Bush insists,]$_{Attr1}$ [it would be a great tool for curbing the budget deficit and slicing the lard out of government programs.]$_{Arg1}$ [as a result]$_{conn}$[He **wants** it now.]$_{Arg2}$ WSJ_0609
Expansion	Ex. (3.1) ["**The Red Cross has been** helping people for 125 years.]$_{Arg1}$ [and]$_{conn}$[**New York Life has been** doing the same for over 140 years" .]$_{Arg2}$ WSJ_0453
	Ex. (3.2) [**Many** of Nasdaq's biggest technology stocks were in the forefront of the rally.]$_{Arg1}$ [for example]$_{conn}$ [**Microsoft** added 2 1/8 to 81 3/4 and **Oracle Systems** rose 1 1/2 to 23 1/4.]$_{Arg2}$ WSJ_0327
Temporal	Ex. (4.1) [But many of his statements on the issue in Parliament subsequently were proven wrong by documentary evidence.]$_{Arg1}$ [since then]$_{conn}$[The scandal has faded and flared.]$_{Arg2}$ WSJ_2041
	Ex. (4.2) [In 1900, for instance, less than 8% of assets went into bank deposits.]$_{Arg1}$ [then]$_{conn}$[That rose to nearly 18% during the Depression.]$_{Arg2}$ WSJ_1755

Table 1: Implicit discourse relations in the PDTB. Original PDTB annotations are delimited by square brackets and a suffix. We further annotate certain text spans with wavy lines to indicate entity continuity, and **bold** to indicate the lexical evidence for specific semantic characteristics per relation type. The corresponding source file is cited at the end of each instance.

arguments: the opposing argument cannot manifest negation nor negative expressions. This is because either negate–negate or negate–negative interaction in two predicates express agreement instead of contradiction. For example, in (Ex. 2.1) "not"–"ruined" is a negate–negative interaction indicating agreement, implying a *Contingency* relation.

Topic continuity is also deemed important as described in the PDTB annotation guidelines: "Arg1 and Arg2 share a predicate or a property and the difference is highlighted with respect to the values assigned to this property" (Prasad et al. 2007). We observe that the shared topic usually appears as part of the predicate or subject. In (Ex. 1.1), the two arguments share the topic "prevent" which acts as the predicate of Arg2; in (Ex. 1.2), the continued topic, "government", is the subject of Arg1.

3.2 Contingency

A pair of arguments establish a *Contingency* relation when "one argument casually influences the other". When an instance is annotated as *Contingency*, connectives such as *because*, *so*, *therefore* can be inserted between the arguments (Prasad et al. 2007). For example, *Contingency* can be employed to objectively narrate causal facts. For example, juxtaposing "It's raining" and "It is wet on the ground" implies causality between the facts. The recognition of such causality is difficult because it requires world knowledge.

In the PDTB context, *Contingency* also exists in subjective situations. Importantly, these can be identified through surface linguistic clues. The following are two prototypical instances of subjective cases:

Subjective Judgement. A *Contingency* relation is applied when a person gives justification for his subjective judgement or opinion. A strong level of subjectivity in statements often necessitates its justification. In the PDTB, this pattern can manifest when the two arguments are quotations due to a person. In Arg1, the individual gives a strong judgement which begs justification, which is then given in Arg2. For example, Arg1 in (Ex 2.1) is a subjective judgement: "Psyllium's not a good crop", attributed to Sooraji Jath; in Arg2, Mr Jath is quoted with his justification – because it might be ruined by rain at the wrong time.

The semantic characteristic of this case is the strong subjective judgement which is reflected in a "be" + <adjective> structure or similar pattern, where the adjective is usually judgemental; "good" in (Ex. 2.1). This semantic characteristic is a valid signal for *Contingency* in the environment where both arguments originate from the same source, e.g. Sooraji Jath in (Ex. 2.1). This source is annotated as a relation–level attribution span.

The source speaker's subjectivity can also manifest in the lexical choice of the main verb in the attribution span. In (Ex. 2.2), the Carl Schramm's "scoff[ing]"requires justification, hence the presence of Arg2.

Intention. The subjectivity of *Contingency* relations can also relate to an agent's intention. We find two surface patterns that capture intent, based on either statement or action:

i) *IntentionSay*: An agent has an intention in one argument and states their reason in the opposing argument. For example, Arg2 in (Ex. 2.4) states President Bush's intention, and Arg1 is his quotation to explain why "he wants it now".

ii) *IntentionDo*: An agent has an intention in one argument which motivates their action in the opposing argument. This intention is usually used to support the action.

In most realizations in the PDTB, intention appears in the predicate of one argument, thus requiring topic continuity or attribution in the other argument to license the *Contingency* relation. For example, *IntentionSay* requires that the agent with the intention has topic continuity with the subject of the attribution of the opposing argument; similarly *intentionDo* requires the agent to have topic continuity with the subject in the opposing argument.

3.3 Expansion

Expansion covers those relations which "expand the discourse and move its narrative or exposition forward" (Prasad et al. 2007). The semantic characteristic of *Expansion* is thus the notion of topic continuity itself. Our analysis finds that the most representative topic continuity pattern for *Expansion* is the "General(G)–Specific(S)" pattern. In this pattern, a general statement starts a text, which is then followed by specific statement(s) that give detail.

The toy example of "Many IT companies are near Silicon Valley (G). Google is in Mountain View (S1). Facebook is in Menlo Park (S2).", demonstrates two topic continuity patterns: i) narrowing topic continuity (G→S1, G→S2) and ii) parallel continuity (S1→S2). Both are common realizations in the PDTB of this "General–Specific" *Expansion* pattern.

i) *Narrowing Entity Continuity:* Arg1 exhibits an umbrella concept, usually accompanied by an indefinite pronoun in subject part or predicate part, such as "everyone" in (Ex. 3.2). Arg2 then gives a specific hyponym/meronym that illustrates Arg1, often in the form of an actual named entity in subject.

ii) *Parallel Entity Continuity:* This case usually has the topic continuity between entities of the same conceptual level, commonly in the form of similar entities as the subject. For example, "The Red Cross" and "New York Life" in (Ex. 3.1) are both organizational entities. Besides, two arguments' predicates also have topic continuity, e.g. "has been" in (Ex. 3.1).

While the "General–Specific" patterns appears often, we note that many instances are missed by current automated systems (inclusive of ours) due to the difficulty of recognizing possible manifestations of topic continuity, which is discussed later.

3.4 Temporal

Temporal relations "[describe situations where] the arguments are related temporally" (Prasad et al. 2007). We find *Temporal* relation are particularly difficult to distinguish, as temporal relationships are universal, and an intrinsic property of events, inclusive of those annotated with the other three top-level discourse relations. Since temporal relations always hold between two events (especially for those causally related events where the reason is described before the result), implicit temporal relations are hardly marked by characteristic patterns.

For example, in the *Contingency* relation of (Ex. 2.3), Arg1 of "Mr. Sim set growth ..." occurs before its Arg2 "He took the company ...", and could be construed to take on a *Temporal* relation. In an opposing line of reasoning, the *Temporal* relation of (Ex. 4.1), "statement ... were proven wrong" could be construed for *Contingency* as it causes the "scandal has faded and flared". In (Ex. 4.2), one might infer a *Comparison* relation because Arg1 indicates that the "bank deposit" is small while Arg2 says it "rose". We posit that temporal relations often hold when other discourse relations lack compelling evidence; hence we deem *Temporal* as a default classification.

While unappealing, this characteristic is not problematic in the PDTB: instances tagged as *Temporal* relation in PDTB are infrequent, accounting for 5% of the total implicit discourse instances in PDTB. For these reasons, we limit our focus to the other three implicit discourse relations and leave the exploration of *Temporal* to future work.

4 Quantitative Methodology

Our approach is to implement a feature-based supervised baseline that is comparable to the current state-of-the-art and augment it with designed features, motivated by the previous corpus study. We adopt this approach (rather than induce features from deep neural networks), as it allows more introspection and transparency of the final model.

We propose complex features leveraging various lexicons to capture such linguistic properties. Note that different semantic characteristics manifest different ways of cohesion with respect to topic continuity and attribution. To this end, we design a bespoke set of features for each semantic characteristic, and another set of features for their corresponding cohesion device. As we have seen that their joint presence licenses the discourse relations, we compose new features by exhaustively pairing the semantic characteristic feature set with ones modeling the two topical cohesion devices.

As most key information appears in the argument's subject or predicate, we only attempt to capture information related to either the subject or predicate. Here, we use "subject" to refer the subject of the main verb and its modifiers, as determined by the Stanford Dependency Tree convention (De Marneffe and Manning 2008) and use "predicate" to refer the main verb, its modifiers and its complements: `xcomp`, `ccomp`.

We now describe our features to recognize topic continuity, and then describe features engineered to model the relation-specific characteristics. We finish this section by discussing feature selection to reduce noise.

4.1 Topic Continuity

We handle two basic cases for topical cohesion: (co-)reference and repetition; other cases are mentioned in Section 6. To recognize references, we employ the Stanford Coreference Resolution System (Clark and Manning 2016). It can recognize antecedent–anaphor pairs such as "President Bush ↔ He" in (Ex. 2.4). However, it often misses repetitions such as the "prevent–prevent" pair in (Ex. 1.1). To address this shortcoming, we add two handcrafted rules to capture repetition topic continuity manifestations:

i) The repeated word is an open class word and acts at least in one argument as the subject or predicate.

ii) The auxiliaries of the main verbs in both arguments are identical.

If a pair of repeated words from two arguments satisfy either criterion, we treat the words as topic continuity.

4.2 Feature Engineering for Semantic Characteristics

To recap, we proposed features specifically to capture negation, subjectivity and intention, and narrowing and parallel entity continuity as these are key characteristics for *Comparison*, *Contingency* and *Expansion*.

Negation. This phenomenon is modeled by three feature sets: `NegFS` for negation, `NegCoheFS` for its corresponding cohesion devices, and a composite feature set `NegFS`⊗`NegCoheFS` to capture their coordination. `NegFS` consists of three binary features {`Arg1Neg`, `Arg2Neg`, `bothNeg`}; where the individual features indicate where the negative expression is found. For example, (Ex. 1.1) would generate `Arg2Neg`, capturing the presence of "not" in Arg2, and the explicit lack of a negate nor negative expression in Arg1. A negative expression is detected when a word is connected by a `neg` edge in the parsed dependency tree as obtained from the Stanford Parser (Manning et al. 2014); specifically, a negative expression is detected as long as a word has a negative tag in the General Inquirer lexicon (Stone, Dunphy, and Smith 1966).

`NegCoheFS` contains an inventory of 8 binary features. As discussed in Section 3.1, topic continuity for negation must appear in at least the subject or predicate position of an argument. Features in `NegCoheFS` capture the topic continuity between an argument's predicate or subject position and its opposing argument (for the opposing argument, we forgo modeling the position of the cohesive topic). For instance, as (Ex. 1.1) has the repetition of the word "prevent" between Arg2's predicate and Arg1, our system generates the feature `Arg2PrediRep`. By similar means, we further define `Arg2SubjRep`, `Arg1SubjRep` and `Arg1PrediRep` for the remaining three repetition patterns. In this way, we have four features for repetition. In the same manner, we define four features for (co-)reference, respectively.

Subjective Judgment is similarly captured with three feature sets: `SubjtivFS` for subjectivity, `SubjtiveCoheFS` for corresponding cohesion devices, and `SubjtivFS`⊗`SubjtiveCoheFS` for their joint presence. `SubjtivFS` comprises of `Arg1Subjtiv`, `Arg2Subjtiv` and `bothSubjtiv`. As per our observations (cf. Sec. 3.2), subjective judgment can be realized in a copular pattern paired with an adjective. To capture this pattern, we infer the feature `Arg1Subjtiv` when the main clause of only Arg1 follows the pattern "be" + <adjective> and either the subject or predicate contains a subjective word as tagged by the Multi-Perspective Question and Answer Corpus (MPQA) (Wilson, Wiebe, and Hoffmann 2005) or a superlative adjective.

The cohesion device feature set `SubjtiveCoheFS` contains 6 binary features, encoding the existence of Arg1's, Arg2's and relation's attribution span (`HasAttr1`, `HasAttr2`, `HasRelAttr` separately) and also encoding whether each attribution span contains strong subjective words (`Attr1Subjtiv`, `Attr2Subjtiv`, `RelAttrSubjtiv`, respectively).

Intention is also similarly captured in three feature sets: `intentFS`, `IntentCoheFS` and their composition `intentFS`⊗`IntentCoheFS`. Here, `intentFS` consists of 3 binary features of `Arg1Intent`, `Arg2Intent` and `bothIntent` captures the location of the intention expressions. For example, if intent is detected only in Arg1's predicate part (not in Arg2's predicate part), the feature `Arg1Intent` will be active ("objective" in Ex. 2.3).

However, in contrast to negation and subjectivity, there is no existing lexicon for intention words. Inspired by (Pontiki and Papageorgiou 2014), we generate our own intention lexicon. We start with basic intention words as seeds, i.e. *want, will, purpose, plan, intend, goal,* and *eager*. We then obtain all the seeds' synonyms from Roget's Thesaurus (Jarmasz 2012) as candidates. Finally, we calculate a confidence score as defined in Eq. (1) for each candidate and prune candidates below the threshold, retaining word w if $c_w > \lambda$ where $\lambda = 0.4$ is set through cross validation on the training set. In Eq. (1), $Cont(w)$ means the *Contingency* instance set that contains the candidate intention word w. Similarly, $intentionDo(i)$ and $intentionSay(w)$ means the instance set that contains *intentionSay* and *intentionDo* patterns discussed in Section 3.2 separately. Here, the constant 1 is used for smoothing.

$$c_w = \frac{\#Cont.(w) \cap (intentionDo(w) \cup intentionSay(w)) + 1}{\#intentionDo(w) \cup intentionSay(w) + 1} \quad (1)$$

`IntentCoheFS` includes 6 features to model the presence of the three possible location of topic continuity – in both arguments, between Arg1 and Attr2, or between Arg2 and Attr1 – via the two devices of repetition and (co-)reference. For example in (Ex. 2.3), the subjects of the two arguments "Mr. Sim" and "He" are recognized as a co-reference pair by Stanford Coreference Resolution system. Therefore, (Ex. 2.3) has `Arg1SubjArg2Subj2-Coref`. Similarly, (Ex. 2.4) has a co-reference between Attr1 and Arg2's subject, thus the feature `Attr1SubjArg2Subj-Coref` is active.

Narrowing Entity Continuity. We use a single feature `NarrowingConti` to encode this phenomenon. Indefinite pronouns (e.g. "everyone" in Arg1 of (Ex. 3.2)) is tagged as "InDEF" and "DEF4" in the General Inquirer. Named Entities are further recognized by Stanford CoreNLP (Manning et al. 2014). If an instance has an indefinite pronoun in the subject or the predicate of Arg1 and has a recognized named entity as the subject of Arg2, the feature `NarrowingConti` will be active.

Parallel Entity Continuity. Similarly, one feature handles this pattern. If an instance has the same type of the named entity as the subject and possesses any form (i.e., either reference and repetition) topic continuity in the predicate, we activate the feature `ParaConti`.

4.3 Feature Selection

The resultant system has many features and can thus overfit. To reduce noise, we select relation-specific features through cross validation on the training set. For each relation, we rank all features in the full feature set S, from least significant to most significant, by their χ^2 association with the relation type. We then iteratively remove features one at a time, when the feature is found unhelpful in cross validation in binary relation classification. Using this method, we obtain four feature sets, i.e. $S_{Comp.}$, $S_{Cont.}$, $S_{Exp.}$, $S_{Temp.}$.

For binary classification, we use feature set specifically selected for one relation. For four-way classification, we take the union of all the four sets of features as the final feature set for classification.

5 Experiments

We evaluate our linguistic observations by testing our features' effectiveness in one-versus-other binary and four-way classification. Following previous work (Pitler, Louis, and Nenkova 2009; Park and Cardie 2012; Rutherford and Xue 2014), we adopt the Naïve Bayes classifier. We adopt the standard PDTB v2.0 dataset partitioning, using S2-20, S0-1 and S23-24, and S21-22 for training, development and testing, respectively, and follow the practice in (Zhou et al. 2010; Liu and Li 2016; Lei et al. 2017) to admit a separate category for Entity Relations (*EntRel*; hence distinct from *Expansion*).

We implemented features in (Rutherford and Xue 2014) as the baseline. These include Brown clustering pairs, production rules and context, General Inquirer tags, verb classes and sentiment and polarity. We remove features which appear less than 5 times. Following (Rutherford and Xue 2014), we use all instances with re-weighting.

As the number of features in our proposed scheme (less than 100 in total) are much fewer than those of the baseline (over 160K), we adopt a stacked structure to allow our features a fair chance of influencing the decision boundaries. We use the respective (one-vs-other or four-way) baseline system to generate a prediction, which is then used as an input feature to the final classifier.

Model	Comp.	Cont.	Exp.	Temp.	4-way
1. (Ji and Eisenstein 2015)	35.93	52.78	–	27.63	–
2. (Liu and Li 2016)	37.91	55.88	69.97	37.17	44.98
3. (Chen et al. 2016)	40.17	54.76	–	31.32	–
4. (Qin, Zhang, and Zhao 2016)	41.55	57.32	71.50	35.43	–
5. (Liu et al. 2016)	39.86	54.48	70.43	38.84	46.29
6. (Qin et al. 2017)	40.87	54.56	72.38	36.20	–
7. (Lei et al. 2017)	40.47	55.36	69.50	35.34	46.46
8. (Rutherford and Xue 2014)	39.70	54.42	70.23	28.69	–
9. Baseline	38.41	53.88	72.22	27.46	44.93
10. All features	**43.24**	**57.82**	72.88	29.10	**47.15**

Table 2: F_1 comparison among existing models and our model "All features". Statistically significant results ($p < 0.05$) over the baseline are bolded.

The main experimental results (Table 2, Row 10 vs. Rows 1–7) show that our system achieves a new state-of-the-art performance level in all tasks with the exception of the *Temporal* relation. Our system builds on our re-implemented baseline (Row 9), which is itself comparable in results with the original paper (Rutherford and Xue 2014; Row 8). As we did not design features specifically for the *Temporal* relation, it is perhaps unsurprising that we do not outperform in this subtask; however, *Temporal* demonstrates gains over the baseline, since our proposed features still help to lessen the number of false positives.

Feature set	Comp.	Cont.	Exp.	Temp.	4-way
All features	43.24	57.82	72.88	29.10	47.15
w/o Negation	-4.68	0.0	0.0	-0.07	-0.45
w/o Intention	-0.86	-2.29	0.0	0.0	-1.32
w/o Subjective	-0.97	-1.39	-0.05	-1.52	-0.24
w/o Parallel Entity	-0.09	-0.18	-0.15	0.0	-0.3
w/o Narrowing Entity	-0.09	-0.46	-0.05	0.0	-0.26

Table 3: Relative F_1 performance with feature ablation. "All features" denotes our complete system.

We next conducted feature ablation experiments that largely confirmed the impact of the each individual patterns (Table 3). We see a sharp decrease for *Comparison* when removing `Negation` features, and similarly for *Contingency* when eliminating `Intention` and `Subjective` feature sets. Such drops show how those features are integral to capturing the corresponding relations. However, `Parallel Entity` and `Narrowing Entity` feature sets have less impact on *Expansion*. Our error analyses uncovered two reasons: the coverage of those two feature sets is low, and that many instances captured by those two features have been already correctly classified as *Expansion* relation by the baseline features (the recall for baseline *Expansion* binary classification is around 85%). Interestingly, the `subjective` pattern helps to recognize *Temporal* relation. By analyzing instances, we found *Temporal* instances often narrates two event objectively, such that subjectivity helps to eliminate false positives.

To gain further understanding, we analyze the top features as selected by χ^2 feature selection. Table 4 lists these features in descending rank, with its correlation as calculated using Pearson correlation. In general, we observe most of the top-ranked features are composite features, most involving a semantic characteristic with one of the two cohesion devices (topic continuity and attribution).

Specifically for *Comparison*, the top four informative features are related to `Arg2Neg`. This corroborates with people's actual strategy in discourse: if we want express negativity, we often start the non-negative clause first (i.e. Arg1 in PDTB). For *Contingency*, both Cont.1 and Cont.3 concern subjectivity, while Cont.2 and Cont.4 relate to the two intention patterns of *intentionDo* and *intentionSay*. Moreover, Cont.5 is designed for *Expansion* which is a strong indicator to prune off false positives for *Contingency* during classification. As for *Expansion*, four negative indicators are ranked highly which again indicate the features we de-

ID	Feature Description	Correlation
Comp.1	Arg2Neg⊗Arg2Subj-Coref	Positive
Comp.2	Arg2Neg⊗Arg2Predi-Rep	Positive
Comp.3	Arg2Neg⊗Arg1Predi-Rep	Positive
Comp.4	Arg2Neg⊗Arg2Subj-Rep	Positive
Comp.5	Arg1Neg⊗Arg1Subj-Rep	Positive
Cont.1	Arg1Subjtiv⊗RelAttr	Positive
Cont.2	Arg2Intent⊗Arg1SubjArg2Subj2-Coref	Positive
Cont.3	Attr1Subjtiv	Positive
Cont.4	Arg2Intent⊗Attr2SubjArg1Subj-Coref	Positive
Cont.5	ParaConti	Negative
Exp.1	Arg2Neg⊗Arg2Subj-Coref	Negative
Exp.2	Arg2Intent⊗Arg1SubjArg2Subj2-Coref	Negative
Exp.3	Arg2Intent⊗Attr2SubjArg1Subj-Coref	Negative
Exp.4	ParaConti	Positive
Exp.5	Arg2Neg⊗Arg2Subj-Rep	Negative
Exp.6	NarrowingConti	Positive
Temp.1	RelAttr	Negative
Temp.2	Arg2Neg⊗Arg2Subj-Coref	Negative
Temp.3	Arg2Neg⊗Arg2Predi-Rep	Negative
Temp.4	Attr1Subjtiv	Negative
Temp.5	Arg1Subjtiv⊗RelAttr	Negative

Table 4: Most important features by χ^2 feature selection per relation type. "⊗" denotes a composite feature constructed from atomic features.

sign have a strong specification for one relation. Moreover, our topic continuity features are also among the top ranked informative features (Exp.4 and Exp.6). Finally, we obtain only negatively correlated features after feature selection for *Temporal*. Interestingly, RelAttr is the most informative feature for *Temporal* relation. By studying its distribution in training data, we find that relation attribution span appears significantly less often than other relations (3% *Temporal* instances contain attribution spans, in contrast with 11% for the other three relation types). This distribution indicates that quoted arguments are dispreferred in expressing *Temporal* relation in PDTB.

6 Discussion: Improving Topic Continuity

We have demonstrated that topic continuity is a key source of knowledge that informs the discourse relation between arguments. Improving our automated methods for capturing topic continuity will thus plough returns into implicit discourse recognition. We discuss two directions for improving the current results.

The first direction is to address errors originating from the existing components, such as the coreference resolution system. We employed the Stanford Coreference Resolution System (Clark and Manning 2016) due to its simplicity of integration. It achieves an F_1 score of 65.29%, which is fairly low. Potentially better results may be achieved by integrating stronger coreference libraries. A failure case is illustated in Ex. 1, where ideally "does" should be label as co-referent with "includes money spent on residential renovation":

1. The government includes money spent on residential renovation;]$_{Arg1}$[in contrast]$_{conn}$ [Dodge does**n't**.]$_{Arg2}$ WSJ_0036 (*Comparison*)

A second direction is to capture other realizations of topic continuity beyond repetition and reference. Although both reference and repetition are common, we acknowledge that traditional systemic functional linguistics – *e.g.*, Halliday (1976) – specifies other devices for topic continuity: inclusive of Lexical, Substitution and even Ellipsis (*ibid.*).

From our observations, the Lexical device is the most frequent cohesion tie in the PDTB. Authors often use words with a certain relationship (antonyms, hypernyms, etc.) to imply reference to the same topic, hence achieving topic continuity.

2. Drug companies in the key index also notched gains.]$_{Arg1}$ [for instance]$_{conn}$ [Wellcome gained 18 to 666 on a modest 1.1 million shares.]$_{Arg2}$ WSJ_0137 (*Expansion*)

3. Wellcome gained 18 to 666 on a modest 1.1 million shares.]$_{Arg1}$ [and]$_{conn}$[Glaxo, the U.K.'s largest pharmaceutical concern, advanced 23 to #14.13.]$_{Arg2}$ WSJ_0137 (*Expansion*)

Ex. 2 illustrates narrowing entity continuity. Here, "companies" and "Wellcome" is a hypernym–hyponym pair forming a Lexical cohesion tie. Capturing such pairs requires significant real-world knowledge and still difficult with the current state-of-the-art in NLP. Ex. 3 shows parallel continuity. While both arguments have the same type of named entity as their subject, their predicates are synonymous: "gain" and "advance". We note that the lexical cohesion tie is most representative manifestation of both narrowing entity and parallel entity continuity patterns for *Expansion*. This qualitatively provides another reason why our specifically-designed Parallel Entity and Narrowing Entity features for *Expansion* are of limited help in Table 3.

7 Conclusion

We conducted a comprehensive analysis on the implicit discourse relations in the Penn Discourse Treebank through a corpus study. Our study discovered representative, relation-specific linguistic properties for the *Comparison*, *Contingency* and *Expansion* relation types, but finds little specific evidence for the *Temporal* relation.

We find that discourse relations are licensed by cooperation between two linguistic properties: i) a discourse relation's unique semantic characteristics and ii) a cohesion device: either topic continuity and attribution. We capture these intuitions by engineering feature sets for these two properties, and their combinations. When used in a Naïve Bayes classifier, we achieve a new level of state-of-the-art performance, bettering strong baselines inclusive of recent neural models.

We further discussed how topic continuity can manifest in a variety of manners, some of which is beyond what is currently computationally feasible to accurately capture. This paints one future direction forward as computational methods for detecting lexical cohesion improve beyond simple repetition and (co-)reference.

References

Agirre, E.; Banea, C.; Cer, D. M.; Diab, M. T.; Gonzalez-Agirre, A.; Mihalcea, R.; Rigau, G.; and Wiebe, J. 2016.

Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval@ NAACL-HLT*, 497–511.

Carlson, L.; Okurowski, M. E.; and Marcu, D. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Chen, J.; Zhang, Q.; Liu, P.; Qiu, X.; and Huang, X. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Clark, K., and Manning, C. D. 2016. Improving coreference resolution by learning entity-level distributed representations. *ACL*.

De Marneffe, M.-C., and Manning, C. D. 2008. Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.

Jarmasz, M. 2012. Roget's thesaurus as a lexical resource for natural language processing. *arXiv preprint arXiv:1204.0140*.

Ji, Y., and Eisenstein, J. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics* 3:329–344.

Lei, W.; Wang, X.; Liu, M.; Ilievski, I.; He, X.; and Kan, M.-Y. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. *IJCAI*.

Lin, Z.; Kan, M.-Y.; and Ng, H. T. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 343–351. Association for Computational Linguistics.

Liu, Y., and Li, S. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. *EMNLP*.

Liu, Y.; Li, S.; Zhang, X.; and Sui, Z. 2016. Implicit discourse relation classification via multi-task neural networks. *AAAI*.

Louis, A.; Joshi, A.; Prasad, R.; and Nenkova, A. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 59–62. Association for Computational Linguistics.

MacCartney, B. 2009. *Natural language inference*. Stanford University.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, 55–60.

Marcu, D., and Echihabi, A. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 368–375. Association for Computational Linguistics.

Park, J., and Cardie, C. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 108–112. Association for Computational Linguistics.

Pitler, E.; Louis, A.; and Nenkova, A. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 683–691. Association for Computational Linguistics.

Pontiki, M., and Papageorgiou, H. 2014. theres no way i would ever buy any mp3 player with a measly 4gb of storage: Mining intention insights about future actions. In *International Conference on HCI in Business*, 233–244. Springer.

Prasad, R.; Miltsakaki, E.; Dinesh, N.; Lee, A.; Joshi, A.; Robaldo, L.; and Webber, B. L. 2007. The penn discourse treebank 2.0 annotation manual.

Qin, L.; Zhang, Z.; Zhao, H.; Hu, Z.; and Xing, E. P. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. *ACL*.

Qin, L.; Zhang, Z.; and Zhao, H. 2016. A stacking gated neural architecture for implicit discourse relation classification. *EMNLP*.

Rutherford, A., and Xue, N. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *EACL*, volume 645, 2014.

Stone, P. J.; Dunphy, D. C.; and Smith, M. S. 1966. The general inquirer: A computer approach to content analysis.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347–354. Association for Computational Linguistics.

Zhou, Z.-M.; Xu, Y.; Niu, Z.-Y.; Lan, M.; Su, J.; and Tan, C. L. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 1507–1514. Association for Computational Linguistics.