Adjacency Pair Recognition in Wikipedia Discussions using Lexical Pairs

Emily K. Jamison ‡ **and Iryna Gurevych** †‡
‡Ubiquitous Knowledge Processing Lab (UKP-TUDA),
Department of Computer Science, Technische Universität Darmstadt
† Ubiquitous Knowledge Processing Lab (UKP-DIPF),
German Institute for Educational Research
http://www.ukp.tu-darmstadt.de

Abstract

Adjacency pair recognition, a necessary component of discussion thread reconstruction, is the task of recognizing reply-to relations between pairs of discussion turns. Previously, dialogue act classification and metadata-based features have been shown useful in adjacency pair recognition. However, for certain forums such as Wikipedia discussions, metadata is not available, and existing dialogue act typologies are inapplicable. In this work, we show that adjacency pair recognition can be performed using lexical pair features, without a dialogue act typology or metadata, and that this is robust to controlling for topic bias of the discussions.

1 Introduction

A growing cache of online information is contained inside user-posted forum discussions. Thread structure of the discussion is useful in extracting information from threads: Wang et al. (2013) use thread structure to improve IR over threads, and Cong et al. (2008) use thread structure to extract question-answer pairs from forums. However, as Seo et al. (2009) point out, thread structure is unavailable in many forums, partly due to the popularity of forum software phpBB[1] and vBulletin[2], whose default view is non-threaded.

Thread reconstruction provides thread structure to forum discussions whose original thread structure is nonexistant or malformed, by sorting and reordering turns into a directed graph of *adjacency* (reply-to) relations. Pairs of adjacent turns (*adjacency pairs*) were first identified by Sacks et al. (1974) as the structural foundation of a discussion, and recognition of adjacency pairs is a critical step in thread reconstruction (Balali et al., 2014; Wang et al., 2008; Aumayr et al., 2011).

Turn1: *This article has been gutted. I deleted a lot of the cruft that had taken over, but a lot of former material is missing.[...]*

Turn2: *Good; the further this nest of doctrinaire obscurities is gutted, the better.*

Turn3: *Wait, you changed it to say that English doesn't have a future tense or you're citing that as an error (which it would naturally be)? For what it matters, [...]*

Turn4: *English doesn't have a future tense. It indicates the future with a modal (will) used with the present-tense inflection of the verb. [...]*

Figure 1: Excerpt from the EWDC discussion *Grammatical Tense:gutted.*

Figure 1 shows an excerpt from Ferschke's (2014) English Wikipedia Discussions Corpus. Thread structure is indicated by tab indents. Turn pairs *(1,2)*, *(1,3)*, and *(3,4)* are adjacency pairs; pairs *(2,3)* and *(1,4)* are not. Adjacency pair recognition is the classification of a pair of turns as adjacent or nonadjacent.

Although most previous work on thread reconstruction takes advantage of metadata such as user id, timestamp, and quoted material (Aumayr et al., 2011; Wang et al., 2011a), metadata is unreliable in some forums, such as Wikipedia Discussion page forums, where metadata and user contribution is difficult to align (Ferschke et al., 2012). Wang et al. (2011b) find that joint prediction of dialogue act labels and adjacency pair recognition improves accuracy when compared to separate classification; dialogue act classification does not require metadata. However, existing dialogue act typologies are unapplicable for some forums (see Section 2.2).

In this paper, we perform adjacency pair recognition on pairs of turns extracted from the English

[1] http://www.phpbb.com/
[2] http://www.vbulletin.com/

28th Pacific Asia Conference on Language, Information and Computation pages 479–488

Wikipedia Discussions Corpus (*EWDC*). We use lexical pair features, which require neither metadata nor development of a dialogue act typology appropriate for Wikipedia discussions. We perform two sets of supervised learner experiments. First, we use lexical pairs for adjacency pair recognition in K-fold Cross Validation (*CV*) setting. Then we show how this permits topic bias, inflating results. Second, we repeat our first set of experiments, but in a special CV setting that removes topic bias. We find that lexical pairs outperform a cosine similarity baseline and a most frequent class baseline both without and with controlling for topic bias, and also exceed the performance of lexical strings of stopwords and discourse connectives on the task.

2 Background

Adjacency pairs were proposed as a theoretical foundation of discourse structure by Sacks et al. (1974), who observed that conversations are structured in a manner where the current speaker uses structural techniques to select the next speaker, and this structure is the adjacency pair: a pair of adjacent discussion turns, each from different speakers, and the relation between them.

2.1 Adjacency Pair Typologies

Previous work on adjacency pair recognition has found adjancency pair typologies to be useful (Wang et al., 2011b). Early work on adjacency pair typologies labelled adjacency pairs by adjacency relation function. Schegloff and Sacks (1973) proposed initial sequences (e.g., greeting exchanges), preclosings, pre-topic closing offerings, and ending sequences (i.e., terminal exchanges). Other adjacency pair typologies consist of pairs of dialogue act labels. Based on their work with transcripts of phone conversations, Sacks et al. (1974) suggested a few types of adjacency pairs: greeting-greeting, invitation-acceptance/decline, complaint-denial, compliment-rejection, challenge-rejection, request-grant, offer-accept/reject, question-answer. In transcribed phone dialogues on topics of appointment scheduling, travel planning, and remote PC maintenance, Midgley et al. (2009) identified adjacency pair labels as frequently co-occurring pairs of dialog acts, including suggest-accept, bye-bye, request/clarify-clarify, suggest-reject, etc.

2.2 Discussion Structure Variation

Much adjacency pair descriptive work was based on transcriptions of phone conversations. Sacks et al. (1974) were discussing phone conversations when they observed that a speaker can select the next speaker by the use of adjacency pairs, and the subsequent speaker is obligated to give a response appropriate to and limited by the adjacency pair, such as answering a question. In a phone conversation, the participant set is fixed, and rules of the conversation permit the speaker to address other participants directly, and obligate a response.

However, in other types of discussion, such as forum discussions, this is not the case. For example, in QA-style forums such as CNET (Kim et al., 2010), a user posts a question, and anyone in the community may respond; the user cannot select a certain participant as the next speaker. Wikipedia discussions vary even further from phone conversations: many threads are initiated by users interested in determining community opinion on a topic, who avoid asking direct questions. Wikipedia turns that might have required direct replies from a particular participant in a speaker-selecting (*SS*) phone conversation, are formulated to reduce or remove obligation of response in this non-speaker-selecting context. Some examples are below; NSS turns are actual turns from the EWDC.

Rephrasing a user-directed command as a general statement:

SS turn: "Please don't edit this article, because you don't understand the concepts."

NSS turn: "Sorry, but anyone who argues that a language doesn't express tense [...] obviously doesn't understand the concept of tense enough to be editing an article on it."

Obtaining opinions by describing past user action instead of questioning:

SS turn: "Which parts of this article should we delete?"

NSS turn: "This article has been gutted. I deleted a lot [...]."

Using a proposal instead of a question:

SS turn: "Should we rename this article?"

NSS turn: "I propose renaming this article to [...]"

Following questions with statements that deflect need for the question to be answered:

NSS turn: "Wait, you changed it to say that English doesn't have a future tense or you're citing that as an error (which it would naturally be)? For what it matters, even with the changes, this entire article needs a rewrite from scratch because so much of it is wrong."

Avoiding questions to introduce a new topic:

SS turn: "Have you heard of Flickr?"

NSS turn: "I don't know whether you know about Flickr or not, but theres a bunch of creative commons licensed images here some better and some worse than the article which you might find useful[...]".

Anticipating responses:

NSS turn: "What are the image names? :Image:Palazzo Monac.jpg has a problem, it's licensed with "no derivative works" which won't work on Commons.[...] If you meant other ones, let me know their names, ok?"

As seen above, Wikipedia discussions have different dialogue structure than phone conversations. Because of the different dialogue structure, existing adjacency pair typologies developed for phone conversations are not appropriate for Wikipedia discussions. As it would require much effort to develop an appropriate adjacency-pair typology for Wikipedia discussions, our research investigates the cheaper alternative of using lexical pairs to recognize adjacency pairs.

3 Related Work

To the best of our knowledge, our work is the first work that uses lexical pairs to recognize adjacency pairs.

3.1 Adjacency Pair Recognition

Most previous work on thread reconstruction has, in addition to using metadata-based features, used word similarity, such as cosine similarity or semantic lexical chaining, between turn pairs for adjacency pair recognition or thread structure graph construction. Wang and Rosé (2010) trained a ranking classifier to identify "initiation-response" pairs consisting of quoted material and the responding text in Usenet `alt.politics.usa` messages, based on text similarity features (cosine, LSA). Aumayr et al. (2011) reconstructed discussion thread graphs using cosine similarity between pairs of turns, as well as reply distance, time difference, quotes, and thread length. They first learned a pairwise classification model over a class-balanced set of turn pairs, and then used the predicted classifications to construct graphs of the thread structure of discussions from the Irish forum site `Boards.ie`. Wang et al. (2011a) also reconstructed thread graphs using cosine similarity in addition to features based on turn position, timestamps, and authorship, using forum discussions from Apple Discussion, Google Earth, and CNET. Wang et al. (2008) reconstructed discussion threads of player chats from the educational legislative game *LegSim*, using TF-IDF vector space model similarity between pairs of turns to build the graphs. Balali et al. (2014) included a feature of TF-IDF vector-space model of text similarity between a turn and a combined text of all comments, a feature of text similarity between pairs of turns, and an authorship language model similarity feature, to learn a pairwise ranking classifier, and then constructed graphs of the thread structures of news forum discussions. Wang et al. (2011c) evaluated the use of WordNet, Roget's Thesaurus, and WORDSPACE SemanticVector lexical chainers for detecting semantic similarity between two turns and their titles, to identify thread-linking structure. Wang et al. (2011b) used a dependency parser, based on unweighted cosine similarity of titles and turn contents, as well as authorship and structural features, to learn a model for joint classification of Dialogue Acts and "inter-post links" between posts in the CNET forum dataset.

3.2 Lexical Pairs

We use lexical pairs as features for adjacency pair recognition. Although not previously been used for this task, lexical pairs have been helpful for other discourse structure tasks such as recognising discourse relations. Marcu and Echihabi (2002) used lexical pairs from all words, nouns, verbs, and cue-phrases, to recognise discourse relations. A binary relation/non-relation classifier achieves 0.64 to 0.76 accuracy against a 0.50 baseline, over approx. 1M instances. Lin et al. (2009) performed discourse relation recognition using lexical pairs as well as con-

stituent and dependency information of relations in the Penn Discourse Treebank. They achieved 0.328 accuracy against a 0.261 most frequent class baseline, using 13,366 instances. Pitler et al. (2009) performed binary discourse relation prediction using lexical pairs, verb information, and linguistically-motivated features, and achieve improvements of up to 0.60-0.62 accuracy, compared with a 0.50 baseline, on datasets sized 1,460 to 12,712 instances from the Penn Discourse Treebank. Biran and McKeown (2013) aggregated lexical pairs as clusters, to combat the feature sparsity problem. While improvements are modest, lexical pairs are helpful in these discourse tasks where useful linguistically-motivated features have proven elusive.

4 Dataset

Our dataset[3] consists of discussion turn pairs from Ferschke's (2014) English Wikipedia Discussions Corpus (EWDC). Discussion pages provide a forum for users to discuss edits to a Wikipedia article.

We derived a class-balanced dataset of 2684[4] pairs of adjacent and non-adjacent discussion turn pairs from the EWDC. The pairs came from 550 discussions within 83 Wikipedia articles. The average number of discussions per article was 6.6. The average number of extracted pairs per discussion was 4.9. The average turn contained 81±95 tokens (standard deviation) and 4±4 sentences. To reduce noise, usernames and time stamps have been replaced with generic strings.

4.1 Indentation Reliability

Adjacency is indicated in the EWDC by the user via tab indent, as can be seen in Figure 1.

Incorrect indentation (i.e., indentation that implies a reply-to relation with the wrong post) is quite common in longer discussions in the EWDC. In an analysis of 5 random threads longer than 10 turns each, shown in Table 1, we found that 29 of 74 total turns, or 39%±14pp of an average thread, had indentation that misidentified the turn to which they were a reply. We also found that the misindentation existed in both directions: an approximately equal number of tabs and tab deletions were needed in each article to correct the misindented turns.

Discussion	# Turns	% Misind.	R	L	P(pos)
Grammatical_tense	20	.50	8	7	10/10
Hurricane_Iniki:1	15	.2	2	4	2/3
Hurricane_Iniki:2	13	.46	11	4	5/7
Possessive_adjective	13	.23	1	5	9/10
Prince's_Palace_of_Monaco	13	.54	9	9	6/6
Average	14.8	.39	6.2	5.8	.89

Table 1: Analysis of wrong indentation in 5 discussions, showing misindentation rate, the sum of how many tabs to the left or right are needed to fix the misindented response turn, and P of extracted positive pairs.

To minimize the number of turn pairs with incorrect indentation extracted from the corpus, we extracted our positive and negative pairs as follows: An adjacent pair is defined as a pair of turns in which one turn appears directly below the other in the text, and the latter turn is indented once beyond the previous turn. A non-adjacent pair is defined as a pair of turns in which the latter turn has fewer indents than the previous turn. Our extraction method yields 32 true positives and 4 false positives (precision = 0.89) in the 5 discussions. Analysis of 20 different pairs in Section 7.2 yielded 0.90 class-averaged precision.

5 Human Performance

We annotated a subset of out data, to determine a human upper bound for adjacency pair recognition. Two annotators classified 128 potential adjacency pairs (23 positive, 105 negative) in 4 threads with an average length of 6 turns. The annotators could see all other turns in the conversation, unordered, along with the pair in question. This pairwise binary classification scenario matches the pairwise binary classification in the experiments in Sections 7 and 9. Each pair was decided independently of other pairs. Cohen's kappa agreement (Cohen, 1960) between the annotators was 0.63.

We noticed a common pattern of disagreement in two particular situations. When an "I agree" turn referred back to an adjacency pair in which one turn elaborated on the other, it was difficult for an annotator to determine which member of the original adjacency pair was the parent of the "I agree" comment. In a different situation, sometimes a participant contributed a substantially off-topic post that spawned a new discussion. It was difficult for the annotators to determine whether the off-topic post

[3] www.ukp.tu-darmstadt.de/data/wikidiscourse

[4] Lexical pairs use a large feature space, and dataset size was constrained by computational feasability.

was a vague response to an existing post, or whether the off-topic post was truly the beginning of a brand-new discussion, albeit using the same original discussion thread.

6 Features

We use three types of features for adjacency pair recognition: *lexical pairs*, *structural context information*, and *pair symmetry*.[5]

Lexical pairs. A lexical pair feature consists of a pair of ngrams with one ngram taken from the first document and one ngram taken from the second document. An ngram is a string of consecutive tokens of length n in a text. Following Marcu and Echihabi (2002), we find a relation (in our case, adjacency) that holds between two text spans, N_1, N_2, is determined by the ngram pairs in the cartesian product defined over the words in the two text spans $(n_i, n_j) \in N_1 \times N_2$.

The goal of using lexical pairs is to identify word pairs indicative of adjacency, such as (*why, because*) and (*?, yes*). These pairs cannot be identified using text similarity techniques used in previous work (Wang and Rosé, 2010).

In addition to lexical pairs created from document ngrams, lexical pairs were created from a list of 50 stopwords (Stamatatos, 2011), Penn Discourse Treebank discourse connectives (Prasad et al., 2008), and a particularly effective combination of just 3 stopwords: *and*, *as*, *for*. Other variables included the parameter ngram `n`, and removed stopwords, which skipped unallowed words in the text.

Structural context information. Some of our feature groups include structural context information of the discussion turn codified as lexical items in the lexical pair string. We include sentence boundaries (SB), commas (CA), and sentence location (i.e., sentence occurs in first quarter, last quarter, or middle of the discussion turn). A sample lexical string representing text from the beginning of a turn is below.

Text: *No, that is correct.*
Lexical string: `no-that-is-correct`
with struct.: `no-CA-that-is-correct-SBBEGIN`

Pair symmetry. Our dataset of discussion turn pairs retains the original order from the discussion. This permits us to detect order-sensitive features such as *(why, because)* and not *(because, why)*, in which the ngram from Turn1 always occurs on the left-hand side of the feature name. Adjacency pairs, by definition, are nonsymmetrical. To confirm this property, in some of our feature groups, we extract a reverse-ordered feature for each standard feature. An example with symmetrical and non-symmetrical features is shown below.

Turn1: *Why ?*
Turn2: *Because .*
Non-Sym features: `(why, because)`
Sym features: `(why, because),(because, why)`

7 Experiments without Topic Bias Control

In our first set of experiments, we perform adjacency pair recognition without topic bias control ("non-TBC"). We use the SVM classifier SMO (Hall et al., 2009) in the DKPro TC framework (Daxenberger et al., 2014) for pairwise classification[6] and 5-fold[7] cross-validation (*CV*), in which all instances are randomly assigned to CV folds. These experiments do not control for any topic bias in the data. Previous work (Wang and Rosé, 2010) has structured adjacency pair recognition as a ranking task, with the classifier choosing between one correct and one incorrect response to a given turn. In our experiments, we use pairwise binary classification, because the high indentation error rate and our EWDC instance selection method did not yield enough matched turn pairs for ranking. Feature parameters (such as top k ngrams, string lengths, and feature combinations) were tuned using CV on a development subset of 552 pairs, while the final results reflect experiments on the remaining dataset of 2684 pairs. Results are shown as F-measure

[5] Because our goal is adjacency pair recognition based on text content features, we do not use indentation offset as a feature.

[6] Although discourse turns are sequential, we classify individual pairs. Future work may investigate this as a sequence labelling task.

[7] Although 10-fold CV is more common in many NLP experiments, we use 5-fold cross validation (*CV*) in Section 7 to make our results directly comparable with results in Section 9.

for class c=adjacent, nonadjacent): $F_{1_c}=\frac{2 \cdot P_c \cdot R_c}{P_c+R_c}$, and Accuracy=$\frac{TP+TN}{TP+FP+TN+FN}$. The most frequent class (`MFC`) baseline chooses the most frequent class observed in the training data, as calculated directly from the experiment. The cosine similarity (`CosineSim`) baseline is an SVM classifier trained over cosine similarity scores of the turn pairs. The `Human Upper Bound` shows agreement from Section 5 and reflects a natural limit on task performance.

7.1 Results

Table 2 shows our feature combinations and results. All experiment combinations were $p \leq 0.05$ significantly different (McNemar, 1947) from the `CosineSim` and `MFC` baselines. The highest performing feature combination was pair unigrams with stopwords removed (`pair1grams+noSW`), which had higher accuracy (.68±.02) than all other feature combinations, including `pair1grams` that included stopwords (.64±.01), and all of the stopword feature sets. Stopword removal increases the system performance for our task, which is unexpected because in other work on different discourse relation tasks, the removal of stopwords from lexical pairs has hurt system performance (Blair-Goldensohn et al., 2007; Marcu and Echihabi, 2002; Biran and McKeown, 2013).

Longer ngrams did not increase performance: `pair2grams` (.57±.03) significantly underperformed `pair1grams` (.64±.01).

We examined the performance curve using various n numbers of most frequent lexical pairs as features on a subset of our corpus (1,380 instances). We found that there was no sharp benefit from a few particularly useful pairs, but that performance continued to increase as n approached 5000.

We found that the classifier performs better when the model learns turn pair order, and the reduced data sparsity from using symmetrical features was not valuable (`Stopwords+SB+noSym`, .62 ±.01 versus `Stopwords+SB+Sym`, .55 ±.02). We found that including sentence boundaries was helpful (`Stopwords+SB+noSym`, .60 ±.01 versus `Stopwords+noSB+noSym`, .62 ±.01, significance p=0.05), but that commas and sentence location information were not useful (`Stopwords+SB+CA+SL+noSym`, .61±.01).

Despite their connections with discourse structure, discourse connectives (`DiscConn+SB+noSym`, .61±.01) failed to outperform stopwords (`Stopwords+SB+noSym`, .62 ±.01). This may be due to the rarity of discourse connectives in the discussion turns: Turn pairs have an average of 9.0±8.6 (or 6.5±6.3 if *and* is removed from the list) discourse connectives combined, and 118 different discourse connectives are used.

7.2 Error Analysis

We examined five pairs each of *true positives* (TP), *false negatives* (FN), *false positives* (FP), and *true negatives* (TN), one set of four from each fold of the best performing system, `pair1grams+noSW`. Generally, turns from instances classified negative appeared to be shorter in number of sentences than instances classified positive (shown by pairs of texts: TN (3.2±2.2 and 3.0±3.4); FN (3.0±2.2 and 2.2±1.1); versus, TP (4.8±4.7 and 4.4±3.6); FP (7.6±10.3 and 5.2±2.8)). Two of the 20 had incorrect gold classification based on misindentation.

FP's. One instance is misindented. Four of the five FP's appear to require extensive linguistic analysis to properly determine their non-adjacency. For example, one second turn begins, " 'Linking' just distracts from, but does not solve, the main issue", but linking is not discussed in the earlier turn. To solve this, a system may need to determine keywords, match quotations, or summarize the content of the first turn, to determine whether 'linking' is discussed. In another example, the turns can be respectively summarized as, "here is a reference" and "we need to collectively do X." This pair of summaries is never adjacent. Another FP instance cannot be adjacent to any turn, because it states a fact and concludes "This fact seems to contradict the article, doesn't it?" In the final FP instance, both turns express agreement; they start with "Fair enough." and "Right." respectively. This pattern of sequential positive sentiment among adjacency pairs in this dataset is very rare.

FN's. Among FN's, one pair appears nonsensically unrelated and unsolvable, another is misindented, while another requires difficult-even-for-humans coreference resolution. The other two FN's need extensive linguistic analysis. In the first in-

Name	Words	NGram Length	Context	Symmetry	removed words	F1+	F1-	Acc
Chance								.50
MFC						.44	.54	.49±.01
CosineSim						.62	.49	.56±.01
Human Upper Bound						.70	.93	.89
Stopwords+SB+NoSym	stopwords	1-3	SB	-	-	.61	.63	.62±.01
Stopwords+SB+Sym	stopwords	1-3	SB	Sym	-	.54	.56	.55±.02
Stopwords+noSB+noSym	stopwords	1-3	-	-	-	.57	.63	.60±.01
Stopwords+SB+CA+SL+noSym	stopwords	1-3	SB,CA,SL	-	-	.60	.63	.61±.01
DiscConn+SB+noSym	disc. conn.'s	1-3	SB	-	-	.60	.63	.61±.01
And-as-for	"and", "as", "for"	1-3	-	Sym	-	.63	.39	.54±.03
Pair1grams	all words	1	-	-	-	.62	.66	.64±.01
Pair2grams	all words	2	-	-	-	.60	.53	.57±.03
Pair1ngrams+noDC	all words	1	-	-	disc. conn.'s	.64	.66	.65±.02
pair1ngrams+noSW	all words	1	-	-	stopwords	.66	.70	**.68±.02**

Table 2: Non-TBC adjacency pair recognition feature set descriptions and results. F_1 results are shown by adjacent (+) and nonadjacent (-) classes. Accuracy is shown with cross-validation fold standard deviation. `Human Upper Bound` is calculated on a different dataset, which was also derived from the EWDC.

stance, the first turn begins, "In languages with dynamic scoping, this is not the case,[...]," and the other turn replies, "I'll readily admit that I have little experience with dynamic scoping[...]" This may be solvable with centering theoretic approaches (Guinaudeau and Strube, 2013), which probabilistically model the argument position of multiple sequential mentions of an entity such as "dynamic scoping". The second instance consists of a deep disagreement between the two authors, in which they discuss a number of keywords and topic specific terms, disagree with each other, and make conclusions. This instance may need a combination of a centering theoretic approach, opinion mining, and topic modeling to solve.

7.3 Feature Analysis

We examined the top-ranked features from our most accurate system, `pair1grams+noSW` (accuracy = .66±.01), as determined by Information Gain ranking. Of the five lists of features produced during each of the 5 folds of CV, 12 of the top 20 features were in common between all 5 lists, and 11 of these 12 features contained an ngram referencing "aspirin": (*acid*, *asa* (an abbreviation for acetylsalicylic acid, the generic name for *aspirin*), *aspirin*, *acetylsalicylic*, *name*, *generic*). We explain the likely cause of the topicality in feature importance in Section 8, and run a second set of experiments to control topic bias in Section 9.

8 Topic Bias and Control

In Section 7, we showed that lexical pairs are useful for adjacency pair recognition with random CV fold assignment. However, it is possible that the system's good performance was due not to the lexical pairs, but to information leakage of learning a topic model on instances extracted from a single discussion.

Topic bias is the problem of a machine learner inadvertently learning "hints" from the topics in the texts that would not exist in another experiment addressing the same task. Consider a sample dataset which contains 16 adjacent and 0 nonadjacent pairs from an article on *Aspirin*, and 7 adjacent and 9 nonadjacent pairs from an article on *Wales*. A model trained on this corpus will probably find lexical pair features such as (*?, yes*) and (*why, because*) to be highly predictive. But, lexical pairs containing topic-sensitive words such as *aspirin* and *generic* may also be highly predictive. Such a model is recognizing adjacency by topic. To remove this topic bias, instances from a single article should never occur simultaneously in training and evaluation datasets.

Topic bias is a pervasive problem. Mikros and Argiri (2007) have shown that many features besides ngrams are significantly correlated with topic, including sentence and token length, readability measures, and word length distributions. Topic-controlled corpora have been used for authorship identification (Koppel and Schler, 2003), genre detection (Finn and Kushmerick, 2003), and Wikipedia quality flaw prediction (Ferschke et al., 2013).

The class distribution by discussion in our dataset is shown in Figure 2; imbalance is shown by the percentage of positive pairs minus the percentage of negative pairs. Only 39 of 550 discussions contributed an approximately equal number of positive

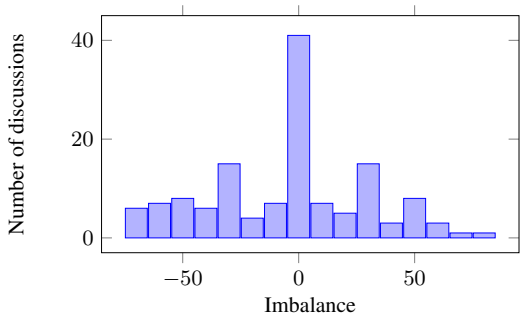

Figure 2: Class imbalance by discussion, in percent. -20 means a discussion is 20 percentile points more negative instances than positive; i.e., if there are 10 instances, 4 positive and 6 negative, then the discussion is a -20 discussion.

and negative instances. 12 discussions contributed only negative instances, and 321 discussions contributed only positive instances[8]. Of discussions with some instances from each class, a whopping 43 of 137 discussions contributed a set of instances that was class imbalanced by 40 percentage points or more. As a result, a classifier will perform above chance if it assumes all instances from one discussion have the same class.

9 Experiments with Topic Bias Control

In our second set of experiments, we performed adjacency pair recognition while controlling for topic bias. To control topic bias, instances from any discussion in a single Wikipedia article are never split across a training and test set. When the cross-validation folds are created, instead of randomly assigning each *instance* to a fold, we assign each *set* of instances from an entire article to a fold. With this technique, any topic bias learned by the classifer will fail to benefit the classifier during the evaluation. We did not use stratified cross-validation, due to the computational complexity of constructing folds of variable-sized threads containing variable class-balance.

We compare against the actual MFC baseline, as seen by the classifier in the experiment. The classifier will perform at this baseline if lexical pairs are not useful for the task. We also compare against cosine similarity, similarly to our previous experiments. The *nonpair 1grams* baseline uses an SVM classifier trained over 5000 individual unigrams from the turn pairs.

[8]Many of these discussions may have consisted of only 2 turns.

Feature	Acc w/o TBC	Acc w TBC
MFC	.49±.01	.44±.04
CosineSim	.56±.01	.54±.06
Nonpair1grams	.67±.02	.49±.03
Stopwords+SB+noSym	.62±.01	.51±.01
Stopwords+SB+Sym	.55±.02	.51±.01
Stopwords+noSB+noSym	.60±.01	.53±.02
Stopwords+SB+CA+SL+noSym	.61±.01	.52±.02
DiscConn+SB+noSym	.61±.01	.51±.02
And-as-for	.54±.03	.49±.03
Pair1grams	.64±.01	**.56±.03**
Pair2grams	.57±.03	.52±.03
Pair1ngrams+noDC	.65±.02	**.56±.03**
Pair1ngrams+noSW	**.68±.02**	.52±.03

Table 3: Adjacency pair recognition, without and with topic bias control.

9.1 Results

The results of our topic bias controlled experiments are shown in Table 3. As entropy decreases with more folds, to avoid exaggerating the reduced entropy effect, 5-fold cross-validation is used. All other experiment parameters are as in Section 7.

All experiment combinations were $p \leq 0.05$ significantly different (McNemar, 1947) from the `CosineSim` and `MFC` baselines, except `Stopwords+SB+CA+SL+noSym`, and all were significantly different from the Nonpair1grams baseline. Absolute classifier performance in the topic bias control paradigm drops significantly when compared with results from the non-topic-bias-control paradigm. This indicates that the classifier was relying on topic models for adjacency pair recognition. Not only is the classifier unable to use its learned topic model on the test dataset, but the process of learning topic modeling reduced the learning non-topic-model feature patterns. Even the feature group `And-as-for` drops, illustrating how topic can also be modelled with stopword distribution, even though the stopwords have no apparent semantic connection to the topic.

The benefit of pair ngrams is shown by the significant divergence of performance of `Nonpair 1grams` and `Pair1grams` in the topic bias control paradigm (.49±.03 versus .56±.03, respectively).

However, several feature sets are still significantly effective for adjacency pair recognition. (`Pair1grams`, `Pair1grams+noDC` perform well above the `MFC` baseline, cosine similarity

baseline, and `Nonpair 1grams` baseline. They also outperform the stopword and the discourse connectives feature sets. The shorter ngrams of `Pair1grams` continue to outperform the bigrams in `Pair2grams`, similarly to the experiments without TBC.

Performance of feature sets exceeding the `MFC` baseline indicates that lexical pair features are informative independently of topic bias.

10 Conclusion

Adjacency pair recognition, the task of discovering reply-to relations between pairs of discussion turns, is a necessary component of discussion thread reconstruction. In this paper, we have evaluated the use of lexical pairs for adjacency pair recognition, and we have shown that they are helpful, outperforming cosine similarity. We have further shown that this benefit is robust to topic bias control.

Our error analysis raises intriguing questions for future research, showing that a number of forms of deeper linguistic analysis, such as keyword extraction, turn summarization, and centering theoretic analysis may be necessary to reduce the current error rate in metadata-less adjacency pair recognition.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Center for Advanced Security Research (`www.cased.de`).

References

Erik Aumayr, Jeffrey Chan, and Conor Hayes. 2011. Reconstruction of threaded conversations in online discussion forums. In *Proceedings of the International Conference on Weblogs and Social Media*, pages 26–33, Barcelona, Spain.

A. Balali, H. Faili, and M. Asadpour. 2014. A supervised approach to predict the hierarchical structure of conversation threads for comments. *The Scientific World Journal*, 2014:1–23.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 69–73, Sofia, Bulgaria.

Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. 2007. Building and refining rhetorical-semantic relation models. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 428–435, Rochester, New York.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 467–474, Singapore.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France.

Oliver Ferschke, Iryna Gurevych, and Marc Rittberger. 2013. The impact of topic bias on quality flaw prediction in Wikipedia. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 721–730, Sofia, Bulgaria.

Oliver Ferschke. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. Ph.D. thesis, Technical University of Darmstadt, Darmstadt, Germany.

Aidan Finn and Nicholas Kushmerick. 2003. Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico. Electronic proceedings.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria, August.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational*

Natural Language Learning, pages 192–202, Uppsala, Sweden.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *IJCAI03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72, Acapulco, Mexico.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

T. Daniel Midgley, Shelly Harrison, and Cara MacNish. 2009. Empirical verification of adjacency pairs using dialogue segmentation. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 104–108, London, UK.

George K. Mikros and Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, Amsterdam, Netherlands. Electronic proceedings.

Emily Pitler, Annie Louis, and Ani Nenkova. 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*, pages 683–691, Suntec, Singapore.

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 Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, Morocco.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Jangwon Seo, W. Bruce Croft, and David A. Smith. 2009. Online community search using thread structure. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1907–1910, Hong Kong, China.

Efstathios Stamatatos. 2011. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527.

Yi-Chia Wang and Carolyn P. Rosé. 2010. Making conversational structure explicit: identification of initiation-response pairs within online discussions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 673–676, Los Angeles, California.

Yi-Chia Wang, Mahesh Joshi, William W Cohen, and Carolyn Penstein Rosé. 2008. Recovering implicit thread structure in newsgroup style conversations. In *Proceedings of the International Conference on Weblogs and Social Media*, pages 152–160, Seattle, Washington.

Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011a. Learning online discussion structures by conditional random fields. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 435–444, Beijing, China.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011b. Predicting thread discourse structure over technical web forums. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 13–25, Edinburgh, UK.

Li Wang, Diana McCarthy, and Timothy Baldwin. 2011c. Predicting thread linking structure by lexical chaining. In *Australasian Language Technology Association Workshop 2011*, page 76, Canberra, Australia.

Li Wang, Su Nam Kim, and Timothy Baldwin. 2013. The utility of discourse structure in forum thread retrieval. In *Information Retrieval Technology*, pages 284–295. Springer.