Identifying Causal Relations Using Parallel Wikipedia Articles

Christopher Hidey
Department of Computer Science
Columbia University
New York, NY 10027
chidey@cs.columbia.edu

Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY 10027
kathy@cs.columbia.edu

Abstract

The automatic detection of causal relationships in text is important for natural language understanding. This task has proven to be difficult, however, due to the need for world knowledge and inference. We focus on a sub-task of this problem where an open class set of linguistic markers can provide clues towards understanding causality. Unlike the explicit markers, a closed class, these markers vary significantly in their linguistic forms. We leverage parallel Wikipedia corpora to identify new markers that are variations on known causal phrases, creating a training set via distant supervision. We also train a causal classifier using features from the open class markers and semantic features providing contextual information. The results show that our features provide an 11.05 point absolute increase over the baseline on the task of identifying causality in text.

1 Introduction

The automatic detection of causal relationships in text is an important but difficult problem. The identification of causality is useful for the understanding and description of events. Causal inference may also aid upstream applications such as question answering and text summarization. Knowledge of causal relationships can improve performance in question answering for "why" questions. Summarization of event descriptions can be improved by selecting causally motivated sentences. However, causality is frequently expressed implicitly, which requires world knowledge and inference. Even when causality is explicit, there is a wide variety in how it is expressed.

Causality is one type of relation in the Penn Discourse Tree Bank (PDTB) (Prasad et al, 2008). In general, discourse relations indicate how two text spans are logically connected. In PDTB theory, these discourse relations can be marked explicitly or conveyed implicitly. In the PDTB, there are 102 known explicit discourse markers such as "and", "but", "after", "in contrast", or "in addition". Of these, 28 explicitly mark causal relations (e.g., "because", "as a result", "consequently").

In addition to explicit markers, PDTB researchers recognize the existence of an open class of markers, which they call *AltLex*. There is a tremendous amount of variation in how AltLexes are expressed and so the set of AltLexes is arguably infinite in size. In the PDTB, non-causal AltLexes include "That compares with" and "In any event." Causal AltLexes include "This may help explain why" and "This activity produced."

Discourse relations with explicit discourse markers can be identified with high precision (Pitler and Nenkova, 2009) but they are also relatively rare. Implicit relations are much more common but very difficult to identify. AltLexes fall in the middle; their linguistic variety makes them difficult to identify but their presence improves the identification of causality.

One issue with causality identification is the lack of data. Unsupervised identification on open domain data yields low precision (Do et al, 2011) and while supervised methods on the PDTB have improved (Ji and Eisenstein, 2015), creating enough labeled data is difficult. Here, we present a distant supervision method for causality identification that uses parallel data to identify new causal connectives given a seed set. We train a classifier on this data and self-train to obtain new data. Our novel approach uses AltLexes that were automatically identified using semi-supervised learning over a parallel corpus. Since we do not know

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1424–1433,
Berlin, Germany, August 7-12, 2016.

a priori what these phrases are, we used a monolingual parallel corpus to identify new phrases that are aligned with known causal connectives. As large corpora of this type are rare, we used Simple and English Wikipedia to create one.

Section 2 discusses prior research in causality and discourse. Section 4 describes how we created a new corpus from Wikipedia for causality and extracted a subset of relations with AltLexes. In section 5, we recount the semantic and marker features and how they were incorporated into a classifier for causality. We show that these features improve causal inference by an 11.05 point increase in F-measure over a naive baseline in 6. Finally, we discuss the results and future work in 7.

2 Related Work

Recent work on causality involved a combination of supervised discourse classification with unsupervised metrics such as PMI (Do et al, 2011). They used a minimally supervised approach using integer linear programming to infer causality. Other work focused on specific causal constructions events paired by verb/verb and verb/noun (Riaz and Girju, 2013) (Riaz and Girju, 2014). Their work considered semantic properties of nouns and verbs as well as text-only features.

There has also been significant research into discourse semantics over the past few years. One theory of discourse structure is represented in the PDTB (Prasad et al, 2008). The PDTB represents discourse relationships as connectives between two arguments. Early work with the PDTB (Pitler and Nenkova, 2009) showed that discourse classes with explicit discourse connectives can be identified with high accuracy using a combination of the connective and syntactic features. Further work (Pitler et al, 2009) resulted in the identification of implicit discourse relations using word pair features; this approach extended earlier work using word pairs to identify rhetorical relations (Marcu, 2001) (Blair-Goldensohn et al, 2007). These word pairs were created from text by taking the cross product of words from the Gigaword corpus for explicit causal and contrast relations. Others built on this work by aggregating word pairs for every explicit discourse connective (Biran and McKeown, 2013). They then used the cosine similarity between a prospective relation and these word pairs as a feature. Recently, the first end-to-end discourse parser was completed (Lin et al, 2012). This parser jointly infers both argument spans and relations. The current state-of-the-art discourse relation classifier is a constituent parse recursive neural network with coreference (Ji and Eisenstein, 2015).

Our work is similar to previous work to identify discourse connectives using unsupervised methods (Laali, 2014). In their research, they used the EuroParl parallel corpus to find discourse connectives in French using known English connectives and filtering connectives using patterns. Unlike this effort, we created our own parallel corpus and we determined new English connectives.

Compared to previous work on causality, we focus specifically on causality and the AltLex. The work by Do and Riaz used minimally supervised (Do et al, 2011) or unsupervised (Riaz and Girju, 2013) approaches and a slightly different definition of causality, similar to co-ocurrence. The work of Riaz and Girju (2013) is most similar to our own. We also examine causality as expressed by the author of the text. However, they focus on intra-sentence constructions between noun or verb phrases directly whereas we attempt to examine how the AltLex connectives express causality in context. Lastly, Riaz and Girju used FrameNet and WordNet to identify training instances for causal verb-verb and verb-noun pairs (Riaz and Girju, 2014) whereas we use them as features for an annotated training set. Overall our contributions are a new dataset created using a distant supervision approach and new features for causality identification. One major advantage is that our method requires very little prior knowledge about the data and requires only a small seed set of known connectives.

3 Linguistic Background

One disadvantage of the PDTB is that the marked AltLexes are limited only to discourse relations across sentences. We know that there are additional phrases that indicate causality within sentences but these phrases are neither found in the set of Explicit connectives nor AltLexes. Thus we expand our definition of AltLex to include these markers when they occur within a sentence. Although some phrases or words could be identified by consulting a thesaurus or the Penn Paraphrase Database (Ganitkevitch et al, 2013), we still need the context of the phrase to identify causality.

We hypothesize that there is significant linguis-

tic variety in causal AltLexes. In the set of known explicit connectives there are adjectives ("subsequent"), adverbs ("consequently"), and prepositions and prepositional phrases ("as a result"). We consider that these parts of speech and syntactic classes can be found in AltLexes as well. In addition, verbs and nouns often indicate causality but are not considered explicit connectives.

Some obvious cases of AltLexes are the verbal forms of connectives such as "cause" and "result". In addition to these verbs, there exist other verbs that can occur in causal contexts but are ambiguous. Consider that "make" and "force" can replace "cause" in this context:

> The explosion **made** people evacuate the building.
>
> The explosion **forced** people to evacuate the building.
>
> The explosion **caused** people to evacuate the building.

However, the words can not be substituted in the following sentence:

> The baker **made** a cake.
>
> *The baker **caused** a cake.
>
> *The baker **forced** a cake.

Furthermore, verbs such as "given" may replace additional causal markers:

> It's not surprising he is tired **since** he did not get any sleep.
>
> It's not surprising he is tired **given that** he did not get any sleep.

There are also some phrases with the same structure as partial prepositional phrases like "as a result" or "as a result of", where the pattern is preposition and noun phrase followed by an optional preposition. Some examples of these phrases include "on the basis of," "with the goal of," and "with the idea of."

We may also see phrases that are only causal when ending in a preposition such as "thanks to" or "owing to." "Lead" may only be causal as a part of "lead to" and the same for "develop" versus "develop from." In addition, prepositions can affect the direction of the causality. Comparing "resulting in" versus "resulting from", the preposition determines that the latter is of the "reason" class and the former is of the "result" class.

Ultimately, we want to be able to detect these phrases automatically and determine whether they are a large/small and open/closed class of markers.

4 Data

In order to discover new causal connectives, we can leverage existing information about known causal connectives. It should be the case that if a phrase is a causal AltLex, it will occur in some context as a replacement for at least one known explicit connective. Thus, given a large dataset, we would expect to find some pairs of sentences where the words are very similar except for the connective. This approach requires a parallel corpus to identify new AltLexes. As large English paraphrase corpora are rare, we draw from previous work identifying paraphrase pairs in Wikipedia (Hwang et al, 2015).

The dataset we used was created from the English and Simple Wikipedias from September 11, 2015. We used the software WikiExtractor to convert the XML into plain text. All articles with the same title were paired and any extra articles were ignored. Each article was lemmatized, parsed (both constituent and dependency), and named-entity tagged using the Stanford CoreNLP suite (Manning et al, 2014). We wish to identify paraphrase pairs where one element is in English Wikipedia and one is in Simple Wikipedia. Furthermore, we do not limit these elements to be single sentences because an AltLex can occur within a sentence or across sentences.

Previous work (Hwang et al, 2015) created a score for similarity (WikNet) between English Wikipedia and Simple Wikipedia. Many similarity scores are of the following form comparing sentences W and W':

$$s(W, W') = \frac{1}{Z} \sum_{w \in W} \max_{w' \in W'} \sigma(w, w') idf(w) \quad (1)$$

where $\sigma(w, w')$ is a score[1] between 2 words and Z is a normalizer ensuring the score is between 0 and 1. For their work, they created a score where $\sigma(w, w') = \sigma_{wk}(w, w') + \sigma_{wk}(h, h')\sigma_r(r, r')$. σ_{wk} is a distance function derived from Wiktionary by creating a graph based on words appearing in a definition. h and h' are the governors of w and w' in a dependency parse and r and r' are the relation. Similar sentences should have similar structure and the governors of two words in different sentences should also be similar. σ_r is 0.5 if h and h' have the same relation and 0 otherwise.

For this work, we also include partial matches, as we only need the connective and the immediate

[1]The score is not a metric, as it is not symmetric.

Method	Max F1
WikNet	0.4850
WikNet, $\lambda = 0.75$	0.5981
Doc2Vec	0.6226
Combined	0.6263

Table 1: Paraphrase Results

surrounding context on both sides. If one sentence contains an additional clause, it does not affect whether it contains a connective. Thus, one disadvantage to this score is that when determining whether a sentence is a partial match to a longer sentence or a shorter sentence, the longer sentence will often be higher as there is no penalty for unmatched words between the two elements. We experimented with penalizing content words that do not match any element in the other sentence. The modified score, where W and W' are nouns, verbs, adjectives, or adverbs, is then:

$$s(W, W') = \frac{1}{Z} \sum_{w \in W} \max_{w' \in W'} \sigma(w, w') idf(w) - \lambda(|W' - W| + |W - W'|) \quad (2)$$

We also compared results with a model trained using doc2vec (Le and Mikolov, 2014) on each sentence and sentence pair and identifying paraphrases with their cosine similarity.

As these methods are unsupervised, only a small amount of annotated data is needed to tune the similarity thresholds. Two graduate computer science students annotated a total of 45 Simple/English article pairs. There are 3,891 total sentences in the English articles and 794 total sentences in the Simple Wikipedia articles. Inter-annotator agreement (IAA) was 0.9626, computed on five of the article pairs using Cohen's Kappa. We tune the threshold for each possible score: for doc2vec the cosine similarity and for WikNet the scoring function. We also tune the lambda penalty for WikNet. F1 scores were calculated via grid search over these parameters and the best settings are a combined score using doc2vec and penalized WikNet with $\lambda = 0.75$ where a pair is considered to be a paraphrase if either threshold is greater than 0.69 or 0.65 respectively.

Using the combined score we obtain 187,590 paraphrase pairs. After combining and deduping this dataset with the publicly available dataset released by (Hwang et al, 2015), we obtain 265,627 pairs, about 6 times as large as the PDTB.

In order to use this dataset for training a model to distinguish between causal and non-causal instances, we use the paired data to identify pairs where an explicit connective appears in at least one element of the pair. The explicit connective can appear in a Simple Wikipedia sentence or an English Wikipedia sentence. We then use patterns to find new phrases that align with these connectives in the matching sentence.

Class	Type	Subtype
Temporal		
Contingency	Cause	reason
		result
	Pragmatic cause	
	Condition	
	Pragmatic condition	
Comparison		
Expansion		

Table 2: PDTB Discourse Classes

To identify a set of seed words that unambiguously identify causal and non-causal phrases we examine the PDTB. As seen in Table 2, causal relations fall under the Contingency class and Cause type. We consider connectives from the PDTB that either only or never appear as that type. The connective "because" is the only connective to be almost always a "reason" connective, whereas there are 11 unambiguous connectives for "result", including "accordingly", "as a consequence", "as a result", and "thus". There were many markers that were unambiguously not causal (e.g. "but", "though", "still", "in addition").

In order to label paraphrase data, we use constraints to identify possible AltLexes.[2] We used Moses (Koehn et al, 2007) to train an alignment model on the created paraphrase dataset. Then for every paraphrase pair we identify any connectives that match with any potential AltLexes. Based on our linguistic analysis, we require these phrases to contain at least one content word, which we identify based on part of speech. We also draw on previous work (Pitler and Nenkova, 2009) that used the left and right sibling of a phrase. Therefore, we use the following rules to label new AltLexes:

1. Must be less than 7 words.
2. Must contain at least one content word:
   (a) A non-proper noun
   (b) A non-modal and non-auxiliary verb
   (c) An adjective or adverb
3. Left sibling of the connective must be a noun phrase, verb phrase, or sentence.
4. Right sibling of the connective must be a noun phrase, verb phrase, or sentence.

[2]We do not attempt to label arguments at this point.

5. May not contain a modal or auxilary verb.

Because connectives identify causality between events or agents, we require that each potential connective link 2 events/agents. We define an event or agent as a noun, verb, or an entire sentence. This means that we require the left sibling of the first word in a phrase and the right sibling of the last word in a phrase to be an event, where a sibling is the node at the same level in the constituent parse. We also require the left and right sibling rule for the explicit connectives, but we allow additional non-content words (for example, we would mark "because of" as a connective rather than "because." We then mark the AltLex as causal or not causal.

Given that the paraphrases and word alignments are noisy, we use the syntactic rules to decrease the amount of noise in the data by more precisely determining phrase boundaries. These rules are the same features used by Pitler and Nenkova (2009) for the early work on the PDTB on explicit connectives. These features were successful on the Wall Street Journal and they are applicable for other corpora as well. Also, they are highly indicative of discourse/non-discourse usage so we believe that we are improving on noisy alignments without losing valuable data. In the future, however, we would certainly like to move away from encoding these constraints using a rule-based method and use a machine learning approach to automatically induce rules.

This method yields 72,135 non-causal and 9,190 causal training examples. Although these examples are noisy, the dataset is larger than the PDTB and was derived automatically. There are 35,136 argument pairs in the PDTB marked with one of the 3 relations that implies a discourse connective (Implicit, Explicit, and AltLex), and of these 6,289 are causal. Of the 6,289 causal pairs, 2,099 are explicit and 273 contain an AltLex.

5 Methods

Given training data labeled by this distant supervision technique, we can now treat this problem as a supervised learning problem and create a classifier to identify causality.

We consider two classes of features: features derived from the parallel corpus data and lexical semantic features. The parallel corpus features are created based on where AltLexes are used as paraphrases for causal indicators and in what context. The lexical semantic features use FrameNet, WordNet, and VerbNet to derive features from all the text in the sentence pair. These lexical resources exploit different perspectives on the data in complementary ways.

The parallel corpus features encourage the classifier to select examples with AltLexes that are likely to be causal whereas the lexical semantic features allow the classifier to consider context for disambiguation. In addition to the dataset, the parallel corpus and lexical semantic features are the main contributions of this effort.

5.1 Parallel Corpus Features

We create a subclass of features from the parallel corpus: a KL-divergence score to encourage the identification of phrases that replace causal connectives. Consider the following datapoints and assume that they are aligned in the parallel corpus:

I was late **because of** traffic.

I was late **due to** traffic.

We want both of these examples to have a high score for causality because they are interchangeable causal phrases. Similarly, we want non-causal phrases that are often aligned to have a high score for non-causality.

We define several distributions in order to determine whether an AltLex is likely to replace a known causal or non-causal connective. We consider all aligned phrases, not just ones containing a causal or non-causal connective to attempt to reduce noisy matches. The idea is that non-connective paraphrases will occur often and in other contexts.

The following conditional Bernoulli distributions are calculated for every aligned phrase in the dataset, where w is the phrase, s is the sentence it occurs in, c is "causal" and nc is "not causal":

$$p_1 = p(w_1 \in s_1 | rel(s_1) \in \{c\}, w_1 \notin s_2) \quad (3)$$

$$p_2 = p(w_1 \in s_1 | rel(s_1) \in \{nc\}, w_1 \notin s_2) \quad (4)$$

We compare these two distributions to other distributions with the same word and in a different context (where o represents "other"):

$$q_1 = p(w_1 \in s_1 | rel(s_1) \in \{nc, o\}, w_1 \notin s_2) \quad (5)$$

$$q_2 = p(w_1 \in s_1 | rel(s_1) \in \{c, o\}, w_1 \notin s_2) \quad (6)$$

We then calculate $D_{KL}(p_1||q_1)$ and $D_{KL}(p_2||q_2)$. In order to use KL-divergence as a feature, we multiply the score by $(-1)^{p<q}$ and add a feature for **causal** and one for **non-causal**.

5.2 Lexical Semantic Features

As events are composed of predicates and arguments and these are usually formed by nouns and verbs, we consider using lexical semantic resources that have defined hierarchies for nouns and verbs. We thus use the lexical resources FrameNet, WordNet, and VerbNet as complementary resources from which to derive features. We hypothesize that these semantic features provide context not present in the text; from these we are able to infer causal and anti-causal properties.

FrameNet is a resource for frame semantics, defining how objects and relations interact, and provides an annotated corpus of English sentences. WordNet provides a hierarchy of word senses and we show that the top-level class of verbs is useful for indicating causality. VerbNet provides a more fine-grained approach to verb categorization that complements the views provided by FrameNet and WordNet.

In **FrameNet**, a semantic frame is a conceptual construction describing events or relations and their participants (Ruppenhofer et al, 2010). Frame semantics abstracts away from specific utterances and ordering of words in order to represent events at a higher level. There are over 1,200 semantic frames in FrameNet and some of these can be used as evidence or counter-evidence for causality (Riaz and Girju, 2013). In Riaz's work, they identified 18 frames as causal (e.g. "Purpose", "Internal cause", "Reason", "Trigger").

We use these same frames to create a lexical score based on the FrameNet 1.5 corpus. This corpus contains 170,000 sentences manually annotated with frames. We used a part-of-speech tagged version of the FrameNet corpus and for each word and tag, we count how often it occurs in the span of one of the given frames. We only considered nouns, verbs, adjectives, and adverbs. We then calculate $p_w(c|t)$ and c_{wct}, the probability that a word w is causal given its tag t and its count, respectively. The lexical score of a word i is calculated by using the assigned part-of-speech tag and is given by $CS_i = p_{w_i}(c|t_i) \log c_{w_i c t_i}$. The total score of a sequence of words is then $\sum_{i=0}^{n} CS_i$.

We also took this further and determined what frames are likely to be *anti-causal*. We started with a small set of seed words derived directly from 11 discourse classes (types and subtypes from Table 2), such as "Compare", "Contrast", "Explain", "Concede", and "List". We expanded this list using WordNet synonyms for the seed words. We then extracted every frame associated with their stems in the stemmed FrameNet corpus. These derived frames were manually examined to develop a list of 48 anti-causal frames, including "Statement", "Occasion", "Relative time", "Evidence", and "Explaining the facts".

We create an anti-causal score using the FrameNet corpus just as we did for the causal score. The total anti-causal score of a sequence of words is $\sum_{i=0}^{n} ACS_i$ where $ACS_i = p_{w_i}(a|t_i) \log c_{w_i a t_i}$ for anti-causal probabilities and counts. We split each example into three parts: the text before the AltLex, the AltLex, and the text after. Each section is given a causal score and an anti-causal score. Overall, there are six features derived using FrameNet: causal score and anti-causal score for each part of the example.

In **WordNet**, words are grouped into "synsets," which represent all synonyms of a particular word sense. Each word sense in the WordNet hierarchy has a top-level category based on part of speech (Miller, 1995). Every word sense tagged as noun, verb, adjective, or adverb is categorized. Some examples of categories are "change", "stative", or "communication". We only include the top level because of the polysemous nature of WordNet synsets. We theorize that words having to do with change or state should be causal indicators and words for communication or emotion may be anti-causal indicators.

Similar to the FrameNet features, we split the example into three sections. However, we also consider the dependency parse of the data. We believe that causal relations are between events and agents which are represented by nouns and verbs. Events can also be represented by predicates and their arguments, which is captured by the dependency parse. As the root of a dependency parse is often a verb and sometimes a noun or adjective, we consider the category of the root of a dependency parse and its arguments.

We include a categorical feature indicating the top-level category of the root of each of the three sections, including the AltLex. For both sides of the AltLex, we include the top-level category of all arguments as well. If a noun has no category, we mark it using its named-entity tag. If there is still no tag, we mark the category as "none."

VerbNet VerbNet is a resource devoted to storing information for verbs (Kipper et al, 2000).

In contrast to WordNet, VerbNet provides a more fine-grained description of events while focusing less on polysemy. Some examples of VerbNet classes are "force", "indicate", and "wish". In VerbNet, there are 273 verb classes, and we include their presence as a categorical feature. Similar to WordNet, we use VerbNet categories for three sections of the sentence: the text pre-AltLex, the AltLex, and the text post-AltLex. Unlike WordNet, we only mark the verbs in the AltLex, root, or arguments.

Interaction Finally, we consider interactions between the WordNet and VerbNet features. As previous work (Marcu, 2001) (Biran and McKeown, 2013) used word pairs successfully, we hypothesize that pairs of higher-level categories will improve classification without being penalized as heavily by the sparsity of dealing with individual words. Thus we include interaction features between every categorical feature for the pre-AltLex text and every feature for the post-AltLex text.

In all, we include the following features (L refers to the AltLex, B refers to the text *before* the AltLex and A refers to the text *after* the AltLex):

1. FrameNet causal score for L, B, and A.
2. FrameNet anti-causal score for L, B, and A.
3. WordNet top-level of L.
4. WordNet top-level of the root of B and A.
5. WordNet top-level for arguments of B and A.
6. VerbNet category for verb at the root of L.
7. VerbNet top-level category for any verb in the root of B and A.
8. VerbNet top-level category for any verbs in the arguments of B and A.
9. Categorical interaction features between the features from B and the features from A.

6 Results

We evaluated our methods on two manually annotated test sets. We used one of these test sets for development only. For this set, one graduate computer science student and two students from the English department annotated a set of Wikipedia articles by marking any phrases they considered to indicate a causal relationship and marking the phrase as "reason" or "result." Wikipedia articles from the following categories were chosen as we believe they are more likely to contain causal relationships: science, medicine, disasters, history, television, and film. For each article in this category, both the English and Simple Wikipedia articles were annotated. A total of 12 article pairs were annotated. IAA was computed to be 0.31 on two article pairs using Kripendorff's alpha.

IAA was very low and we also noticed that annotators seemed to miss sentences containing causal connectives. It is easy for an annotator to overlook a causal relation when reading through a large quantity of text. Thus, we created a new task that required labeling a connective as causal or not when provided with the sentence containing the connective. For testing, we used CrowdFlower to annotate the output of the system using this method. We created a balanced test set by annotating 600 examples, where the system labeled 300 as causal and 300 as non-causal. Contributors were limited to the highest level of quality and from English-speaking countries. We required 7 annotators for each data point. The IAA was computed on the qualification task that all annotators were required to complete. There were 15 questions on this task and 410 annotators. On this simplified task, the IAA improved to 0.69.

We also considered evaluating the results on the PDTB but encountered several issues. As the PDTB only has a limited set of explicit intra-sentence connectives marked, this would not show the full strength of our method. Many causal connectives that we discovered are not annotated in the PDTB. Alternatively, we considered evaluating on the AltLexes in the PDTB but these examples are only limited to inter-sentence cases, whereas the vast majority of our automatically annotated training data was for the intra-sentence case. Thus we concluded that any evaluation on the PDTB would require additional annotation. Our goal in this work was to identify new ways in which causality is expressed, unlike the PDTB where annotators were given a list of connectives and asked to determine discourse relations.

We tested our hypothesis by training a binary[3] classifier on our data using the full set of features we just described. We used a linear Support Vector Machine (SVM) classifier (Vapnik, 1998) trained using stochastic gradient descent (SGD) through the sci-kit learn package. (Pedregosa et al, 2011)[4] We used elasticnet to encourage sparsity and tuned the regularization constant α through grid search.

We use two baselines. The first baseline is the

[3]We combine "reason" and "result" into one "causal" class and plan to work on distinguishing between non-causal, reason, and result in the future.

[4]We also considered a logistic regression classifier.

	Accuracy	True Precision	True Recall	True F-measure
Most Common Class	63.50	60.32	82.96	69.85
$CONN$	62.21	78.47	35.64	49.02
LS	67.68	61.98	58.51	60.19
KLD	58.03	**91.17**	19.55	32.20
$LS \cup KLD$	**73.95**	80.63	64.35	71.57
$LS \cup LS_{inter}$	72.99	78.54	64.66	70.93
$KLD \cup LS \cup LS_{inter}$	70.09	76.95	58.99	66.78
$LS \cup KLD \cup CONN$	71.86	70.28	**77.60**	**73.76**
$Bootstrapping_1$	79.26	**77.97**	82.64	80.24
$Bootstrapping_2$	**79.58**	77.29	**84.85**	**80.90**

Table 3: Experimental Results

most common class of each AltLex according to its class in the initial training set. For example, "caused by" is almost always a causal AltLex. A second baseline uses the AltLex itself as a categorical feature and is shown as $CONN$ in Table 3. For comparison, this is the same baseline used in (Pitler and Nenkova, 2009) on the explicit discourse relations in the PDTB. We compare these two baselines to ablated versions of our system. We evaluate on the KLD (KLD) and semantic (LS and LS_{inter}) features described in sections 5 and 5.1. LS consists of features 1-8, all the FrameNet, VerbNet, and WordNet features. LS_{inter} includes only the interaction between categorical features from WordNet and VerbNet.

We calculate accuracy and true precision, recall, and F-measure for the causal class. As seen in Table 3, the best system ($LS \cup KLD \cup CONN$) outperforms the baselines.[5] The lexical semantic features by themselves (LS) are similar to those used by (Riaz and Girju, 2014) although on a different task and with the WordNet and VerbNet features included. Note that the addition of the Altlex words and KL divergence ($LS \cup KLD \cup CONN$) yields an absolute increase in f-measure of 13.57 points over lexical semantic features alone.

6.1 Bootstrapping

Our method for labeling AltLexes lends itself naturally to a bootstrapping approach. As we are using explicit connectives to identify new AltLexes, we can also use these new AltLexes to identify additional ones. We then consider any paraphrase pairs where at least one of the phrases contains one of our newly discovered AltLexes. We also use our classifier to automatically label these new data points and remove any phrases where the classifier did not agree on both elements in the pair. The set of features used were the $KLD \cup LS \cup LS_{inter}$ features as these performed best on the development set. We use early stopping on the development data to identify the point when adding additional data is not worthwhile. The bootstrapping method converges quickly. After 2 iterations we see a decrease in the F-measure of the development data.

The increase in performance on the test data is significant. In Table 3, $Bootstrapping_n$ refers to results after n rounds of bootstrapping. Bootstrapping yields improvement over the supervised method with an absolute gain of 7.14 points.

6.2 Discussion

Of note is that the systems without connectives (combinations of LS, LS_{inter}, and KLD) perform well on the development set without using any lexical features. Using this system enables the discovery of new AltLexes during bootstrapping, as we cannot rely on having a closed class of connectives but need a way of classifying connectives not seen in the initial training set.

Also important is that the Altlex by itself ($CONN$) performs poorly. In comparison, in the task of identifying discourse relations in the PDTB these features yield an 75.33 F-score and 85.85% accuracy in distinguishing between discourse and non-discourse usage (Pitler and Nenkova, 2009) and an accuracy of 93.67% when distinguishing between discourse classes. Although this is a different data set, this shows that identifying causality when there is an open class of connectives is much more difficult. We believe the connective by itself performs poorly because of the wide

[5]These results are statistically significant by a binomial test with $p < 7 * 10^{-6}$.

	True Precision	True Recall	True F-measure
$FrameNet$	67.88	53.14	59.61
$WordNet$	76.92	9.52	16.94
$VerbNet$	38.70	3.80	6.92

Table 4: Semantic Feature Ablation

linguistic variation in these alternative lexicalizations. Many connectives appear only once or not at all in the training set, so the additional features are required to improve performance.

In addition, the "most common class" baseline is a strong baseline. The strength of this performance provides some indication of the quality of the training data, as the majority of the time the connective is very indicative of its class in the held-out test data. However, the the overall accuracy is still much lower than if we use informative features.

The KLD and LS feature sets appear to be complementary. The KLD feature sets have higher precision on a smaller section of the data, whereas the LS system has higher recall overall. These lexical semantic features likely have higher recall because these resources are designed to represent *classes* of words rather than individual words. Some connectives occur very rarely, so it is necessary to generalize the key aspects of the connectives and class-based resources provide this capability.

In order to determine the contribution of each lexical resource, we perform additional feature ablation for each of FrameNet, WordNet, and VerbNet. As seen in Table 4, the lexical semantic resources each contribute uniquely to the classifier. The FrameNet features provide most of the performance of the classifier. The WordNet and VerbNet features, though not strong individually, supply complementary information and improve the overall performance of the LS system (see Table 3) compared to just using FrameNet alone.

Finally, the model ($LS \cup KLD \cup CONN$) correctly identifies some causal relations that neither baseline identifies, such as:

> Language is reduced to simple phrases or even single words, eventually **leading to** complete loss of speech.
>
> Kulap quickly accelerated north, **prompting** the PAGASA to issue their final advisory on the system.

These examples do not contain standard causal connectives and occur infrequently in the data, so the lexical semantic features help to identify them.

After two rounds of bootstrapping, the system is able to recover additional examples that were not found previously, such as:

> When he finally changed back, Buffy stabbed him **in order to** once again save the world.

This connective occurs rarely or not at all in the initial training data and is only recovered because of the improvements in the model.

7 Conclusion

We have shown a method for identifying and classifying phrases that indicate causality. Our method for automatically building a training set for causality is a new contribution. We have shown statistically significant improvement over the naive baseline using semantic and parallel corpus features. The text in the AltLex alone is not sufficient to accurately identify causality. We show that our features are informative by themselves and perform well even on rarely occurring examples.

Ultimately, the focus of this work is to improve detection of causal relations. Thus, we did not evaluate some intermediate steps, such as the quality of the automatically annotated corpus. Our use of distant supervision demonstrates that we can use a large amount of possibly noisy data to develop an accurate classifer. To evaluate on the intermediate step would have required an additional annotation process. In the future, we may improve this step using a machine learning approach.

Although we have focused exclusively on Wikipedia, these methods could be adapted to other domains and languages. Causality is not easily expressed in English using a fixed set of phrases, so we would expect these methods to apply to formal and informal text ranging from news and journals to social media. Linguistic expressions of causality in other languages is another avenue for future research, and it would be interesting to note if other languages have the same variety of expression.

References

Or Biran and Kathleen McKeown. 2013. *Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation*. Proceedings of ACL.

Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. 2007. *Building and Refining Rhetorical-Semantic Relation Models*. Proceedings of NAACL-HLT.

Nathanael Chambers and Dan Jurafsky. 2008. *Unsupervised Learning of Narrative Event Chains*. Stanford University. Stanford, CA 94305.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. *Minimally Supervised Event Causality Identification*. Transactions of ACL, 3:329344.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. *Aligning Sentences from Standard Wikipedia to Simple Wikipedia*. Proceedings of NAACL-HLT.

Yangfeng Ji and Jacob Eisenstein. 2015. *One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations*. Proceedings of EMNLP.

Karin Kipper, Hoa Trang Dan, and Martha Palmer. 2000. *Class-Based Construction of a Verb Lexicon*. American Association for Artifical Intelligence.

Majid Laali and Leila Kosseim. 2014. *Inducing Discourse Connectives from Parallel Texts*. Proceedings of COLING: Technical Papers.

Quoc Le and Tomas Mikolov. 2014. *Distributed Representations of Sentences and Documents*. Proceedings of ICML.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. *A PDTB-Styled End-to-End Discourse Parser*. Department of Computer Science, National University of Singapore.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. *The Stanford CoreNLP Natural Language Processing Toolkit*. Proceedings of ACL: System Demonstrations.

Daniel Marcu and Abdessamad Echihabi 2001. *An Unsupervised Approach to Recognizing Discourse Relations*. Proceedings of ACL.

George A. Miller. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM.

The PDTB Research Group 2008. *The PDTB 2.0. Annotation Manual*. Technical Report IRCS-08-01. Institute for Research in Cognitive Science, University of Pennsylvania.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. *Automatic sense prediction for implicit discourse relations in text* Proceedings of ACL.

Emily Pitler and Ani Nenkova. 2009. *Using Syntax to Disambiguate Explicit Discourse Connectives in Text*. Proceedings of ACL-IJCNLP Short Papers.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0*. Proceedings of LREC.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. *Realization of Discourse Relations by Other Means: Alternative Lexicalizations*. Proceedings of COLING.

Kira Radinsky and Eric Horvitz. 2013. *Mining the Web to Predict Future Events*. Proceedings of WSDM.

Mehwish Riaz and Roxana Girju. 2013. *Toward a Better Understanding of Causality between Verbal Events: Extraction and Analysis of the Causal Power of Verb-Verb Associations*. Proceedings of SIGDIAL.

Mehwish Riaz and Roxana Girju. 2014. *Recognizing Causality in Verb-Noun Pairs via Noun and Verb Semantics*. Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. University of California, Berkeley.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. *PPDB: The Paraphrase Database*. Proceedings of NAACL-HLT.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. 2011. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research Volume 12.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: open source toolkit for statistical machine translation*. Proceedings of ACL: Interactive Poster and Demonstration Sessions.