Identifying Causal Relations Using Parallel Wikipedia Articles

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

Abstract

The automatic detection of causal relationships in text is important for natural language understanding. However, this task has proven to be difficult 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. As there is significant variety in these phrases we take advantage of the neural network architecture, where neural language modeling allows us to model similar phrases and predict more accurately phrases we have never seen before. We train a classifier using latent features generated by a recursive neural network and observed features derived from the open class markers and lexical semantics.

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) (PDTB Research Group, 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 thus, 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 method for causality identification that uses parallel data to identify new causal connectives given a seed set. Our novel approach uses AltLexes that were automatically identified using a parallel corpus. Since we do not know *a priori* what these phrases are, we used a monolingual parallel corpus to identify new phrases that are aligned with known causal connectives.

We train a classifier on a training set derived from the monolingual parallel corpus. This classifier uses a neural language model to derive latent features and uses observed features extracted from the connective and surrounding text.

Section 2 discusses prior research in causality, discourse, and neural networks. In section 4, we describe the model and how the neural latent features and observed features are incorporated into a classifier for causality. We show that these features improve causal inference by more than a 12% absolute increase over a naive baseline in 5.

2 Related Work

Recent work on causality involved a combination of supervised discourse relation classification with unsupervised metrics such as pointwise mutual information (PMI) (Do et al, 2011). They used a minimally supervised approach to do joint inference for causality using integer linear programming. Other new work focused on specific causal constructions events paired by verb/verb and verb/noun (Riaz and Girju, 2013) (Riaz and Girju, 2014). Their approach considers semantic properties of nouns and verbs in addition to 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 (PDTB Research Group, 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. The current state-of-the-art discourse relation classifier is a constituent parse recursive neural network that identifies relations given argument spans (Ji and Eisenstein, 2015).

This discourse relation classifier continues a tradition of neural approaches to language. Early work (Bengio et al, 2003) showed the advantages of modeling words in a continuous space. Whereas the curse of dimensionality makes it exceedingly unlikely that we will see a phrase as it increases in length, the neural approach allows for sentences that are similar in vector space to have similar predictive probabilities. More recently, researchers developed methods to learn representations of words using large corpora, where a neural language model predicts words given their contexts using derived latent features to create word embeddings (Mikolov et al, 2013). They showed that these word embeddings have linear properties that make them useful for analogies among other tasks. Other similar work (Levy and Goldberg, 2014) demonstrated that the contexts of a word need not be limited to linear semantic contexts but may produce functionally different embeddings when using, for example, dependency relations as context. Finally, other work (Levy and Goldberg, 2014) showed that training

word embeddings using the skip-gram model with negative sampling is equivalent to factorizing a shifted PMI matrix between words and contexts.

Furthermore, researchers (Socher et al, 2014) have had success using dependency tree recursive neural networks (DT-RNN). They developed these models initially to describe images using sentences. Others (Iyyer et al, 2014) applied these DT-RNNs to "quiz bowl" question answering, where the goal is to determine the correct answer given questions in the field of literature or history.

Our work combines aspects of the lexical semantics work on causality and neural network language modeling. Unlike other approaches to discourse relation identification, we leverage a parallel corpus to train the model and improve continuous sentence representations. We also use a DT-RNN to model these relations, unlike other approaches using constituent parses. As causality involves actions between agents and events, we believe that the DT-RNN better models this interaction because of the dependency tree structure. Finally, we use this approach to discover new connectives we have never seen before.

Overall our contributions are a new model of causality 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 linguistic 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."

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 Methods

As our goal is to predict, for any word or phrase, whether it is indicative of causality, we model this as a binary logistic regression with both latent and observed features (similar to a mixed effects model).

$$p(y_n = 1|\theta) = \sigma(w^T \cdot h_n + \beta^T \cdot \phi_n) \quad (1)$$

This is the probability that a connective can replace a known discourse connective for causality.

Neural language modeling approaches have shown that discrete methods can be improved upon by using continuous representations of words. These representations do not require observing all combinations or sequences.

Because we are trying to discover new causal connectives, representing these connectives in a continuous space should allow us to generalize better to connectives we have never seen before. Furthermore, many connectives are ambiguous and it is important to consider the context they appear in within the sentence. As there are many possible sequences that could trigger causality, we wish to model the entire sentence to account for local context. Rather than use contexts based on linear word windows, we believe it is important to use syntactic contexts. Causal relations represent actions between events and/or agents which are expressed as interactions between nouns and verbs. As shown in previous research (Levy and Goldberg, 2014), dependency contexts emphasize functional similarities (how words are used) rather than semantic (what they mean).

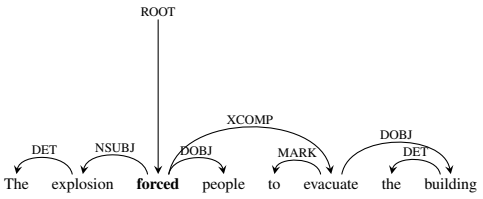

Figure 1: Verbal Causal Connective

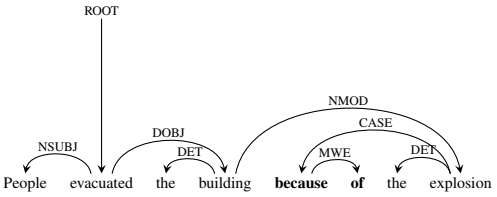

Figure 2: Non-verbal Causal Connective

Thus we model the latent features h using a DT-RNN. We use a similar model to previous work (Iyyer et al, 2014). Each word is represented as a D-dimensional vector. W_e is a $V \times D$ matrix where V is the size of the vocabulary and the embedding of a word w is $W_e(w)$. W_r is a $R \times N \times N$ tensor where R is the number of unique dependency relations and the matrix for relation r is $W_r(r)$. Finally, W_v is a $D \times D$ compositional matrix and b is a D-long vector. We also need to select a non-linear transformation f.

In a DT-RNN, the vector representation of a node is calculated recursively. First, we calculate the value of the leaf nodes from Figure 1:

$$h_{people} = f(W_v^T \cdot W_e(people) + b) \quad (2)$$

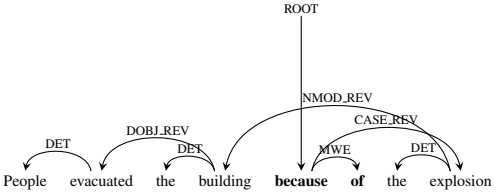

Figure 3: Re-rooted Non-verbal Causal Connective

Then we calculate the value of the parent nodes:

$$\begin{aligned} h_{forced} = f(W_v^T \cdot W_e(forced) + b + \\ W_r(DOBJ)^T \cdot h_{people} + \\ W_r(NSUBJ)^T \cdot h_{explosion} + \\ W_r(XCOMP)^T \cdot h_{evacuate}) \end{aligned} \quad (3)$$

The formula for any node n is then as follows:

$$\begin{aligned} h_n = f(W_v^T \cdot W_e(n) + b + \\ \sum_{k \in K(n)} W_r(Rel(n,k))^T \cdot h_k) \end{aligned} \quad (4)$$

where the parameters to be learned are $\theta = \{W_v, W_e, W_r, b\}$ and f is a non-linearity.

Traditionally, the root of a sentence in a dependency parse is the predicate. Most of the time this is a verb but can be a nominal or adjectival predicate when the main verb is the copula. For a dependency recursive neural network, the hidden state representation of children of each node are combined additively and passed on to their parents. As the gradients vanish the nodes higher in the tree will have more influence. Thus, if possible, the connective should be close to or at the root of the tree.

In the case of a verbal causal connective (Figure 1), the connective is already at the root of the tree. However, if we express the same concept with a non-verbal connective (Figure 2), the connective is no longer the root. We re-write these trees to reverse the path from the original root node to the root node of the connective (Figure 3) This provides for parallel structure between connectives that vary in their part-of-speech. This also allows us to distinguish between multiple instances of causality within a single sentence.

Finally, as all of our training data consists of paraphrases of sentences, we add the following constraint, where s and $s+1$ are paraphrases:

$$\forall_{s,s+1} h(s)^T \cdot h(s+1) \geq \kappa$$

This encourages paraphrases to have similar embeddings.

We do not expect the latent features to fully explain the presence of causality so we create several observed features. We consider three classes of features here: global context features derived using event embeddings, features derived from the parallel corpus data and lexical semantic features.

The global context features are created from the entirety of Wikipedia. We assume that events that are likely to have causal relations will occur in close proximity. Thus we create event embeddings from the dependency trees in a sentence and nearby events.

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.

4.1 Global Context Features

PMI has been shown to be indicative of causality between events (Do et al, 2011). One way to determine PMI would be to take discrete counts of events and their arguments as well as joint counts for pairs of events and arguments. This method would suffer from the curse of dimensionality as we would expect to see some events rarely or not at all.

In order to determine PMI between events, we approach the problem from the opposite perspective. Previous work (Levy and Goldberg, 2014) showed that training word embeddings using a skip-gram model with negative sampling is the same as factorizing a shifted PMI matrix. Thus, given an appropriate event representation, we can create event and context embeddings whose dot product reflects the PMI be-

tween them.

For these event embeddings, we use a scaled-down version of the DT-RNN and consider only a predicate and its arguments. The reason for discarding local context is to have a globally representative embedding of an event that may be adapted according to local context. Using the Stanford dependency parser (Manning et al, 2014), we create collapsed, propagated dependency graphs for all of Wikipedia. We determine predicates by identifying words that have a subject relationship with another word, and then extract any direct objects or indirect objects, if present.

Finally, we require a method for sampling events. Instead of sampling from a joint distribution over events, we instead sample from a factorized distribution to allow for generation of unseen events which are more likely to be a negative sample. We assume the joint distribution of a single event factorizes as follows:

$$p(p, S, O, I) = p(p) \prod_{s \in S} p(s|p) \prod_{o \in O} p(o|p) \prod_{i \in I} p(i|p) \quad (5)$$

where p is the predicate, S is the set of subjects, O is the direct objects, and I is the indirect objects.

We consider only events whose predicate and arguments occur at least 100 times anywhere in Wikipedia. We also limit predicates to verbal ones and disallow nominal and adjectival predicates. as we are interested in events. This results in 75,514,687 events, 25,139,836 of which are unique.

In addition, we create embeddings where the context is limited to where the word "because" is present in the sentence. For each sentence with this connective, we create an event-context pair for the cross-product of the events before and after the connective in both directions. This results in 2,939,254 events, 1,148,301 of which are unique.

4.2 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:

> The building was evacuated **due to** the explosion.
>
> The building was evacuated **because of** the explosion.

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 as well and in other contexts.

The following conditional Bernoulli distributions are calculated for every aligned phrase in the dataset, where w is the phrase and s is the sentence it occurs in:

$$p_1 = p(w_1 \in s_1 | rel(s_1) \in \{causal\}, w_1 \notin s_2) \quad (6)$$

$$p_2 = p(w_1 \in s_1 | rel(s_1) \in \{notcausal\}, w_1 \notin s_2) \quad (7)$$

We compare these two distributions to other distributions with the same word and in a different context:

$$q_1 = p(w_1 \in s_1 | rel(s_1) \in \{notcausal, other\}, w_1 \notin s_2) \quad (8)$$

$$q_2 = p(w_1 \in s_1 | rel(s_1) \in \{causal, other\}, w_1 \notin s_2) \quad (9)$$

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 2 features, one for **causal** and one for **non-causal**.

4.3 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.

These features are described in depth in Appendix B.

5 Results

We trained our model on a large monolingual parallel corpus that was created using a distant supervision approach. A seed set of AltLexes was created using the explicit connectives in the PDTB. New connectives were initially determined as those that aligned with the known connectives. This corpus is described in detail in Appendix A.1.

We evaluated our methods on two manually annotated test sets. We used one of these sets for development and one for testing. For the details of these corpora, see Appendix A.2 and A.3.

5.1 Experimental Details

The matrix W_e was pre-trained using word2vec (Mikolov et al, 2013) with a minimum count of 1 to increase the functional similarity of word embeddings. During training we allowed this matrix to vary because the DT-RNN uses different contexts (syntactic dependencies). Other parameters were initialized by sampling uniformly. The feature weights β were pre-trained on the training set using stochastic gradient descent (SGD) and fixed during training.

D was set to 100 and there are 74 dependency relations in the training set. The loss function was the standard cross-entropy for binary logistic regression. We experimented with both AdaGrad (Duchi et al, 2010) and SGD with the optimal learning rate (Bouttou, 2010). For the non-linearity we used the normalized hyperbolic tangent, as this function worked well in previous work with DT-RNNs (Iyyer et al, 2014).

We used the standard zero-mean Gaussian prior on the parameters W_e, W_v, W_r, and b and determined the regularization parameter λ by evaluating on the development data for the set $\{0.001, 0.0001, 0.00001, 0.000001\}$. We also used elasticnet on β to encourage sparsity with the same range for λ_β.

Because there are more non-causal examples, we balanced the training data by subsampling the non-causal class. We also experimented with oversampling and found that subsampling performed better.

We use two baselines. The first baseline is the most common class of each AltLex according to its class in the training set. A second baseline uses the AltLex and is shown as $CONN$ in Table 1. We compare these two baselines to our system with the full system ($RNN \cup LS \cup KLD$), DT-RNN only (RNN) and observed features only ($LS \cup KLD$).

We also compare the results of the global context features using PMI for both the full (PMI_{FULL}) contexts and causal (PMI_{CAUSAL}) contexts only. We evaluated these as features but they performed poorly so we did not include them.

We calculate accuracy, precision, recall, and F-measure for the causal class. As seen in Table 1, the best systems ($LS \cup KLD \cup RNN$) outperform the baselines. [1] However, there is no statistically significant difference between the top performing systems.

5.2 Analysis

Of note is that the systems with the best observed features and latent features perform equally well. It may be that there are some examples that are "easy" and both systems are able

[1] These results are statistically significant by a binomial test with $p < 1 * 10^{-6}$.

	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
$LS \cup KLD$	79.58	77.29	84.85	80.90
RNN	79.05	75.06	**88.88**	81.39
$LS \cup KLD \cup RNN$	**81.50**	**80.79**	84.12	**82.42**
PMI_{CAUSAL}	51.54	48.29	68.26	56.57
PMI_{FULL}	49.30	46.70	78.00	58.42

Table 1: Experimental Results

to pick up on them. However, both systems perform well without using any n-gram features. This is important for discovering new AltLexes, 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.

It is not necessarily surprising that the full event embeddings perform poorly, as these events do not take discourse into account, and may include contrasting or temporal information. We may also need a more complex model that does not reduce the interaction between events to a scalar. In the future we would like to look into a hierarchical model that incorporates the global and local context.

We examine the differences between the systems with an error analysis. In all examples, the causal connective is in bold text.

First, we consider the true positives, causal examples that were correctly labeled. Here is an example that was found by the RNN model that were not found by the $LS \cup KLD$ model:

> DR6 is highly expressed in the human brain regions most affected by Alzheimer's, **so** it is possible that the N-APP/DR6 pathway might be hijacked in the ageing brain to cause damage.

This example is heavily dependent on syntactic context and features that are more dependent on the connective itself would perform poorly.

This example was found by the $LS \cup KLD$ model and not by the RNN model:

> Consequently, the U.S. could find itself bombing operational missiles were the blockade to fail **to force** Khruschchev to remove the missiles already on the island.

This case may fail in the RNN model because the subjunctive tense causes syntactic issues in the dependency parse, whereas the $LS \cup KLD$ model contains features that depend directly on the connective.

The RNN model has higher recall but lower precision, so we also examine the false positives driving the lower recall.

> The government also planned 2,486 plats in two- **and** four-story buildings in what is called the widow's colony outside Bhopal.

Here the RNN model incorrectly labels the connective as causal when it is clearly a conjunction between adjectives. There are only two other examples containing "and" where this is the case so it may not be a pattern.

We also considered the false negatives, the examples that annotators labeled as causal that no system correctly identified:

> Classical music has many different forms, some of which can be used over a long time span **to make** big compositions.

Word	Positive Predictions	Negative Predictions
impact	51024	2991
let	41263	17025
reaction	38106	0
deal	36062	8969
failure	33956	438
effort	31208	47
increasing	29680	1594
serious	27316	53
massive	27264	0
huge	26827	76
severe	26778	605
reception	25518	211147
responded	26335	2865

Table 2: Discovered Connectives

In this example, there is a causal verb and a nominal event as a result. This example is very difficult for a system to pick up on, as it needs to be able to identify that the event in this case is a noun, when nouns are mostly not events.

Consider also the following example:

> It later reported that it had overestimated the windspeeds **and** lowered the storm's status to a tropical depression.

Identifying this as a causal example requires some world knowledge and inference, that "lower windspeeds" implies a "tropical depression" versus another storm category.

Finally, we use the DT-RNN to predict the use of a word as a causal connective in every sentence in Wikipedia to discover connectives that are not in our training set. We list the top occurring connectives that do not appear at any point in the training set in Table 2.

Some words are definitely causal, such as "responded", "impact" and "reaction." Others such as "effort" make sense in phrases for given a reason such as "in an effort to". Finally, other words are questionable, such as the adjectives "serious," "massive," and "huge." These words would certainly appear often in causal contexts but we would not expect them to be causal in almost all cases.

6 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 a DT-RNN and 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.

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 2013, Sofia, Bulgaria.

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*. Proceedings of EMNLP.

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

Quoc Le and Tomas Mikolov. 2014. *Distributed Representations of Sentences and Documents*. Proceedings of the 31st International Conference on Machine Learning.

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 the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Daniel Marcu and Abdessamad Echihabi 2001. *An Unsupervised Approach to Recognizing Discourse Relations*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

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 and Ani Nenkova. 2009. *Using Syntax to Disambiguate Explicit Discourse Connectives in Text*. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore, 4 August 2009.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. *Automatic sense prediction for implicit discourse relations in text* Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 4 August 2009.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. *Realization of Discourse Relations by Other Means: Alternative Lexicalizations*. Coling 2010: Poster Volume. Beijing, August 2010.

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

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 the SIGDIAL 2013 Conference.

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 the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions.

Yoshua Bengio, Rjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. *A Neural Probabilistic Language Model*. Journal of Machine Learning Research 3 (2003) 11371155.

John Duchi, Elad Hazan, and Yoram Singer. 2010. *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*. Journal of Machine Learning Research, 2011.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. ICLR Workshop, 2013.

Omer Levy and Yoav Goldberg. 2014. *Dependency Based Word Embeddings*. ACL-2014.

Omer Levy and Yoav Goldberg. 2014. *Neural Word Embedding as Implicit Matrix Factorization*. NIPS-2014.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. *Grounded Compositional Semantics for Finding and Describing Images with Sentences*. ACL-2014.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daume. 2014. *A Neural Network for Factoid Question Answering over Paragraphs*. EMNLP-2014.

Leon Bouttou. 2010. *Stochastic Gradient Descent*. http://leon.bottou.org/projects/sgd.

A Corpora

A.1 Training

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':

$$s(W, W') = \frac{1}{Z} \sum_{w \in W} \max_{w' \in W'} \sigma(w, w') idf(w) \tag{10}$$

where $\sigma(w, w')$ is a score between 2 words and Z is a normalizer ensuring the score is between 0 and 1. The score is not a metric, as it is not symmetric. 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. Further details are given in their paper.

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

$$s(W, W') = \frac{1}{Z} \sum_{w \in W} \max_{w' \in W'} \sigma(w, w') idf(w) - \lambda(|W' - W| + |W - W'|) \tag{11}$$

where W and W' are nouns, verbs, adjectives, or adverbs.

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 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

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

Table 3: Paraphrase Results

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 286,356 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.

To identify a set of seed words that unambiguously identify causal and non-causal phrases we examine the PDTB. As seen in table 4, 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 "reason", including "accordingly", "as a consequence", "as a result", and "thus". We also added "so" to the list of result markers, but only when the part of speech is a subordinating conjunction. There were many markers that were unambiguously not causal (e.g. "but", "though", "still", "in addition").

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

Table 4: PDTB Discourse Classes

In order to label paraphrase data, we use constraints to identify possible AltLexes. (We do not attempt to label arguments at this point). 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 of AltLexes, 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 of the following content words:
   (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.
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.

This method yields 72,135 non-causal and 9,190 causal training examples. Although these examples are noisy, the dataset is much 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.

A.2 Development

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. Inter-annotator agreement was computed to be 0.31 on two article pairs using Kripendorff's alpha.

A.3 Testing

Inter-annotator agreement 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. Inter-annotator agreement was computed to be 0.67 using Kripendorff's alpha.

B Lexical Semantic Features

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 4), 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.