Causal Markers across Domains and Genres of Discourse

Rutu Mulkar-Mehta
Information Sciences Institute
University of Southern California
me@rutumulkar.com

Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
gordon@ict.usc.edu

Jerry Hobbs
Information Sciences Institute
University of Southern California
hobbs@isi.edu

Eduard Hovy
Information Sciences Institute
University of Southern California
hovy@isi.edu

ABSTRACT

This paper is a study of causation as it occurs in different domains and genres of discourse. There have been various initiatives to extract causality from discourse using causal markers. However, to our knowledge, none of these approaches have displayed similar results when applied to other styles of discourse. In this study we evaluate the nature of causal markers – specifically causatives, between corpora in different domains and genres of discourse and measure the overlap of causal markers using two metrics – Term Similarity and Causal Precision. We find that causal markers, specially causatives (causal verbs) are extremely domain dependent, and moderately genre dependent.

1. INTRODUCTION

Causality is an important phenomenon in discourse, and plays an important role in NLP tasks of discourse understanding [1] and question answering [4]. Over the years, it has captured the attention of various researchers in NLP and numerous research initiatives [7, 6, 13, 15] have evolved for causal relation extraction. The collective goal of the research community has been to identify and extract causal relations, and various approaches such as supervised learning approaches [4] and heuristics based approaches [7] have been taken. Girju et al. [4], use Wordnet [10] for discovering causal markers and evaluate their system on TREC-9 corpus, giving an overall accuracy of 65%. Other research initiatives such as Blanco et al. [2] have achieve a high precision, but they limit their work to using only two causal markers *because* and *since*, providing no details about the total recall of their system. High recall systems are very important for Learning by Reading [1] systems for the purposes of discourse understanding, and current systems are unable to provide this capability.

Causal relations are usually extracted using causal markers, and the limited growth in the area begs the question of whether there is an inherent relation of causal markers with the domain or genre of discourse which make them difficult to adapt across corpora (please refer to Wolters et al. [18] for further information about domains and genres). Marshman et al. [9] present interesting findings about the portability of causal relation markers in French literature by evaluating the presence of general predefined causal markers across domains and genres of discourse. The authors find that although all the causal markers existed in all the domains, the frequency of occurrence varies largely with the domain selected. There has been similar speculation about adaptability of causal cue words and causatives across domains and genres in English, but there has been no systematic study on causal cue words, or causatives (causal verbs). This paper attempts to address this issue by comparing causal markers across various genres (Newspapers, Blogs, Research Papers) and various domains (Finance, Football, Biomedicine), and evaluating the results on two measures – Term Similarity, Causal Precision. Our work focuses on domain and genre adaptability of causatives (causal verbs [4]). This is the first research initiative in this direction and we aim to uncover the importance of domain and genre modeling for automatic discovery of causal relations.

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.
ACM 978-1-4503-0396-5/11/06.

2. CORPORA: SELECTION AND DETAILS

We selected a collection of four very diverse corpora from three genres (newspapers, blogs, publications) and three domains (finance, football, biomedicine). The details of each corpus are provided below:

1. **Newspaper Articles about Finance:** This corpus is part of the LDC corpus (LDC2005T08) called Discourse GraphBank [16, 17] filtered to contain only Wall Street Journal articles about business and finance. The corpus contains a total of 12157 and 525 sentences. From here on, this corpus will be referred to as *Fin*.
2. **Blog Stories about Football:** This corpus is a subset of blog stories extracted by Gordon et al. [5], and focus on stories describing a game of American football. The corpus contains 9071 words and 568 sentences. From here on, this corpus will be referred to as *Fbl-b*.
3. **Newspaper Articles about Football:** This corpus is part of the LDC - New York Times Annotated corpus (LDC2008T19A), and describes football games. There were a total of 11169 words and 544 sentences in the entire corpus. From here on, this corpus will be referred to as *Fbl-n*.
4. **Scientific Publications about Biomedicine:** This corpus was extracted by Mulkar-Mehta et al. [11], and contains scientific publications from PubMed describing the cell cycle. This corpus contains a total of 6030 words and 155 sentences.

Corpus ID	Kappa
Fbl-n	0.86
Fbl-b	0.94
Fin.	0.85
Bio.	1.0

Table 1: Inter-Annotator Kappa agreement for identifying sentence internal causal relations

We conducted three experiments, evaluating the similarity of causal terms across domains and genre.

1. **Same Genre Different Domains:** *Fbl-n* vs. *Fin*
2. **Different Genres Same Domain:** *Fbl-n* vs. *Fbl-b*
3. **Different Genre Different Domains:** *Fbl-n* vs. *Bio*

The purpose of the experiments was to observe the similarity of causal terms across the dimensions of genre and domain, keeping one variable constant while comparing the other. The results were evaluated on the metrics of term similarity and causal precision, the details of which are elaborated in Section 4.

3. ANNOTATION OF CAUSAL RELATIONS

A subset of sentences in the datasets was independently annotated by 2 annotators. Each annotator was asked to judge whether the given sentence contained a causal relation, and if yes, was asked to mark the causal cue words in the sentence. For instance consider the following sentence from the *Fin* corpus:

Unwilling to put up new money for New Zealand until those debts are repaid, most banks refused even to play administrative roles in the new financing, **forcing** *an embarrassed Nomura to postpone it this week.*

Here '*forcing*' is the causal marker.

The inter annotator agreement was evaluated based on the binary decision of whether a sentence contained a causal connective or otherwise. Table 1 shows the Kappa Agreement [3] for the two annotators in each domain. Scientific publications (*Bio*) and blog stories (*Fbl-n*) had perfect and near perfect agreement scores. The newspaper articles (*Fbl-n* and *Fin*) had a similar inter-annotator agreement showing the similarity in writing style and ambiguous causality mentions in this genre of discourse. The annotations from the primary annotator were taken as the gold standard for evaluation.

Figure 1 shows the total number of causal relations extracted per 1000 words in each of the corpora. The corpus of scientific publications had the most number of causal markers extracted per 1000 words, making it the corpus containing a dense set of causal relations. This is not surprising, as the basic nature of scientific publications is to explain, justify and provide reasons for a phenomenon. The blog stories about football has the sparsest mentions of causality because blog stories mention events happening in a chronological order, often not answering the questions of "how" or "why" certain events happen.

4. RESULTS AND ANALYSIS

We use two evaluation measures to compare the similarity in causal markers in the domains:

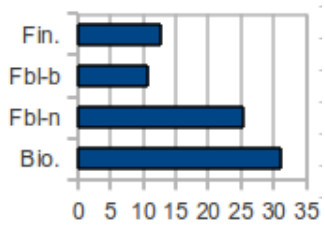

Figure 1: Frequency of Causal Marker per 1000 words

- **Term Similarity:** This is the percentage of overlap in the causation terms between two different corpora. For instance if we have Corpus A and Corpus B, we can use this measure to judge the maximum possible percentage of causal relations that can be extracted from Corpus B, if we are provided with causal markers from Corpus A.
- **Causal Precision:** A term conveying causality in a given context, might not convey causality in another context. In order to measure the causal nature of a term independent of the context, we calculate *Causal Precision*, which is the ratio of the total number of times a term indicates causality and the total number of times a term occurs in discourse. Figure 4 shows the total causal precision of all the causal markers and Kappa agreement of each corpus. This figure shows that except for *Fbl-b*, the annotators agree more when there is less ambiguity in the nature of the causal term. The *Bio* corpus has the highest causal precision and perfect Kappa agreement.

4.1 Comparing All Corpora

All the corpora have five causal markers in common: *after, because, by, to, when*. Of these causal markers, *because* and *when* have high causal precision (greater than 50% for all corpora) showing that these usually mean causality in most domains. However, the rest of the causal markers have other meanings besides causality and have a causal frequency less than 20% for all domains. Figure 2 shows the causal precision and Figure 3 shows the frequency per 1000 words for the common causal markers in all domains. We can see the 'to' has the highest frequency of occurrence in most discourse, but a low causal precision, which means that it has other meanings besides 'in-order-to'.

4.2 Football News (Fbl-n) vs. Biomedical Publications (Bio)

Only 11% of the biomedical causal connectives were found in football news, and 12% of the causal connectives in football news were found in biomedical publications. The small overlap can be attributed to the fact that these corpora did not have a domain or a genre in common. The causal markers limited to the biomedical corpus such as *inhibit, activate, induce* are not found in football articles in the newspaper, which contains causatives such as *edging, lifting*. The few common causal markers shared by these domains were *after, because, by, for, lead, produce, to, when*. These contained only two causatives – *lead* and *produce* and the rest being general conjunctions and prepositions, most of which are polysemous in nature. There were only three causal markers with high causal precision that were common to the two corpora: *because, when produce*. All the other common causal markers had a very low causal precision. This means that causal markers from news-

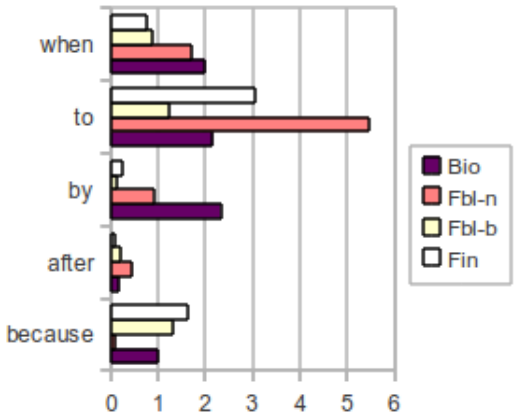

Figure 2: Frequency per 1000 words of the overlapping causal markers between all domains

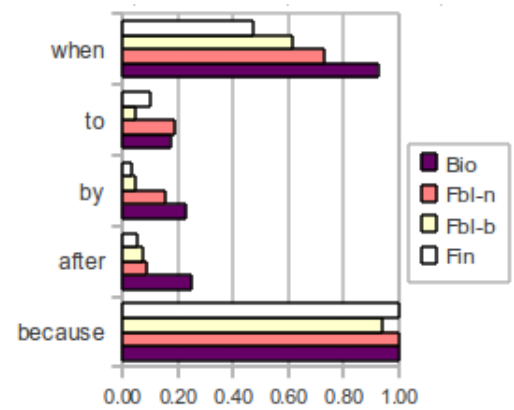

Figure 3: Causal Precision of the overlapping causal markers between all domains

paper articles about football will not help to a large degree for discovering causal connectives in biomedical scientific publications, and vice versa. The obvious explanation is that these domains do not share a similar vocabulary, and hence wont share the same causatives. Table 2 shows the high precision causatives unique to each corpora.

4.3 Football News (Fbl-n) vs. Finance News (Fin)

There was a small increase in term similarity when the domains were selected from the same genre of discourse. Here 22% of the causal markers from finance news were found in football news, and 22% of the causal markers in football news were found in finance news. The differences were attributed to the difference in vocabulary in the two domains, causing different causatives to be used in each domain. For instance in finance articles causatives such as *abolished, stirring, barring* were used, which are rarely ever used in terms of football game descriptions. Similarly causatives such as *lifting, routing* are not present in finance articles. Table 3 shows the high causal precision causatives which were unique to each corpus. The common causal markers between the two domains having a high causal precision were *give, because, help, get* (which are

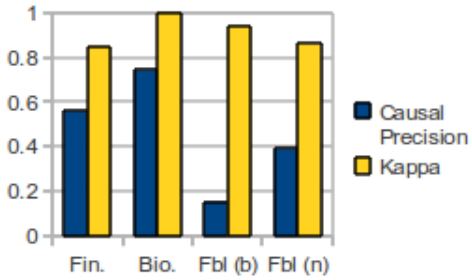

Figure 4: Comparing the percentage of causation terms that convey causality 70% or more times in each corpus, and the Kappa agreement for each corpus

Bio	Fbl-n
promote, control, induce, funnel, govern, trigger, repress, induce, activate, drive, inhibit	snap, subdue, lift, edge, level, lead, hamper, pull, defeat, seal, move, rout, edge, snatch

Table 2: Non-overlapping causatives unique to each domain, in their lemma form

high frequency verbs in newspaper articles), and the rest had much lower causal precision.

4.4 Football News (Fbl-n) vs. Football Blogs (Fbl-b)

The term similarity between two corpora was highest when both the corpora were selected from the same domain. Here 56% of the causal markers that were found in blog stories were also found in the newspaper articles, and 22% of the cause markers found in football news were also found in football blogs. Blog stories were more colloquial as compared to newspaper articles, and used a majority of simple words such as *making, letting* which were not used in newspaper articles, which presented words such as *moving, routing*. Table 4 shows the overlapping causatives in the football news and football blogs. Both the domains shared domain specific causatives, producing a high term similarity between the two corpora.

5. CONCLUSION

In this paper we compare the causal markers, specifically causatives from three domains and three genres of discourse. Our results indicate that there is maximum overlap in causal markers when the corpora share the same domain and least overlap when the corpora do not share either a domain or a genre. In our previous work [12] we were unable to use the domain independent causal markers used in TREC-QA evaluation task by Prager et al. [14] for our task of causality detection, and the causal markers needed to be modeled

Fin	Fbl-n
permit, stir, avert, abolish, elevate, trigger, boost, repeal, raise, rescind, bar, implicate	lift, snap, snatch, rout, produce, halt, roll, put, lift, spark, hampered,

Table 3: Non-overlapping causatives unique to each domain, in their lemma form

Common Causatives	Non Causatives
force, beat, get, give, lead	when, for, by, after, because

Table 4: Common causal markers in *Fbl-n* and *Fbl-b*

specifically for the selected domain. This paper sheds some light on the causes for this, and answers why domain independent causal markers do not provide very good results for causality relation extraction. These findings also justify why causal relations have been so difficult to extract using causal markers, and indicate that some amount of domain understanding is required to obtain high precision and high recall of causal relations. Finally, this work provides the justification for why automated learning techniques have been largely unsuccessful in learning causal relations structures from annotated corpora and applying the learned model to other types of discourse.

Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0172. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, AFRL, ONR, or the US government.

6. REFERENCES

[1] K. Barker, B. Agashe, S.-y. Chaw, J. Fan, N. Friedland, M. Glass, J. R. Hobbs, E. Hovy, D. Israel, D. S. Kim, R. Mulkar-Mehta, S. Patwardhan, B. Porter, D. Tecuci, P. Yeh, and M. Rey. Learning by Reading : A Prototype System , Performance Baseline and Lessons Learned Learning by Reading. *In Proceedings of 22nd National Conference on Artificial Intelligence*, pages 280–286, 2007.

[2] E. Blanco, N. Castell, and D. Moldovan. Causal Relation Extraction. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 310–313, 2008.

[3] J. Cohen. A coefficientof agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[4] R. Girju and D. Moldovan. Mining Answers for Causation. *Proceedings of American Association of Artificial Intelligence*, pages 15–25, 2002.

[5] A. S. Gordon and R. Swanson. Identifying Personal Stories in Millions of Weblog Entries Weblog Stories as Data. *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*, 2009.

[6] T. Inui, K. Inui, and Y. Matsumoto. Acquiring causal knowledge from text using the connective marker tame. *ACM Transactions on Asian Language Information Processing*, 4(4):435–474, Dec. 2005.

[7] C. S. G. Khoo, S. Chan, and Y. Niu. Extracting causal knowledge from a medical database using graphical patterns. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*, pages 336–343, 2000.

[8] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English : The Penn Treebank. *Computational Linguistics*, 1993.

[9] E. Marshman, M.-C. L'Homme, and V. Surtees. Portability of causeâĂŞeffect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora*, 3(2):141–172, Nov. 2008.

[10] G. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, Nov. 1995.

[11] R. Mulkar-Mehta, J. R. Hobbs, C.-C. Liu, and X. J. Zhou. Discovering Causal and Temporal Relations in Biomedical Texts Recognizing Causal and Temporal Relations :. *Proceedings of the AAAI Spring Symposium, Stanford CA*, 2009.

[12] R. Mulkar-Mehta, C. Welty, J. R. Hobbs, and E. Hovy. Using Part-Of Relations for Discovering Causality. *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, 2011.

[13] C. Pechsiri and A. Kawtrakul. Mining Causality from Texts for Question Answering System. *IEICE Transactions on Information and Systems*, E90-D(10):1523–1533, Oct. 2007.

[14] J. M. Prager, J. Chu-Carroll, and K. Czuba. A Multi-Strategy, Multi-Question Approach to Question Answering. *In New Directions in Question-Answering, ed. M. Maybury. Menlo Park, CA: AAAI Press.*, 2004.

[15] B. Rink, C. A. Bejan, and S. Harabagiu. Learning Textual Graph Patterns to Detect Causal Event Relations. *Proceedings of the 23rd Florida Artificial Intelligence Research Society International Conference (FLAIRS'10)*, 2010.

[16] F. Wolf, E. Gibson, A. Fisher, and M. Knight. A procedure for collecting a database of texts annotated with coherence relations. pages 1–23, 2003.

[17] F. Wolf, E. Gibson, A. Fisher, and M. Knight. Discourse Graphbank. *Linguistic Data Consortium, Philadelphia*, 2004.

[18] M. Wolters and M. Kirsten. Exploring the Use of Linguistic Features in Domain and Genre Classification. *Proceedings of EACL 1999*, pages 142–149, 1999.