Causal Markers across Domains and Genres of Discourse

Rutu Mulkar-Mehta, Andrew Gordon, Jerry Hobbs, Eduard Hovy

University of Southern California

me@rutumulkar.com, gordon@ict.usc.edu, hobbs@isi.edu, 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.

Corpora: Selection and Details

1. **Newspaper Articles about Finance:** This corpus is part of the LDC corpus (LDC2005T08) called Discourse GraphBank (Wolf 2003) 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. (Gordon et al. 2009), 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. (Mulkar-Mehta 2009), and contains scientific publications from PubMed describing the cell cycle. This corpus contains a total of 6030 words and 155 sentences.

Experiments

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.

Related Publications

[1] Mulkar, R.; Hobbs, J. R.; and Hovy, E.
Learning from Reading Syntactically Complex Biology Texts.
Proceedings of the AAAI Spring Symposium, Stanford CA, 2007.

[2] Mulkar, R.; Hobbs, J.; Hovy, E.; Chalupsky, H.; and Lin, C.-Y.
Learning by Reading: Two Experiments.
In *Proceedings of 3rd international workshop on Knowledge and Reasoning for Answering Questions*, 2007.

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

[4] 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.

[5] Mulkar-Mehta, R.; Hobbs, J. R.; and Hovy, E.
Granularity in Natural Language Discourse.
International Conference on Computational Semantics, Oxford, UK, 2011.

[6] Mulkar-Mehta, R.; Hobbs, J. R.; and Hovy, E.
Applications and Discovery of Granularity Structures in Natural Language Discourse.
Proceedings of the AAAI Spring Symposium, Stanford CA, 2011.

Annotation of Causal Relation

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. 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.

Evaluation Metrics

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

- **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.

Evaluation Results

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

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

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

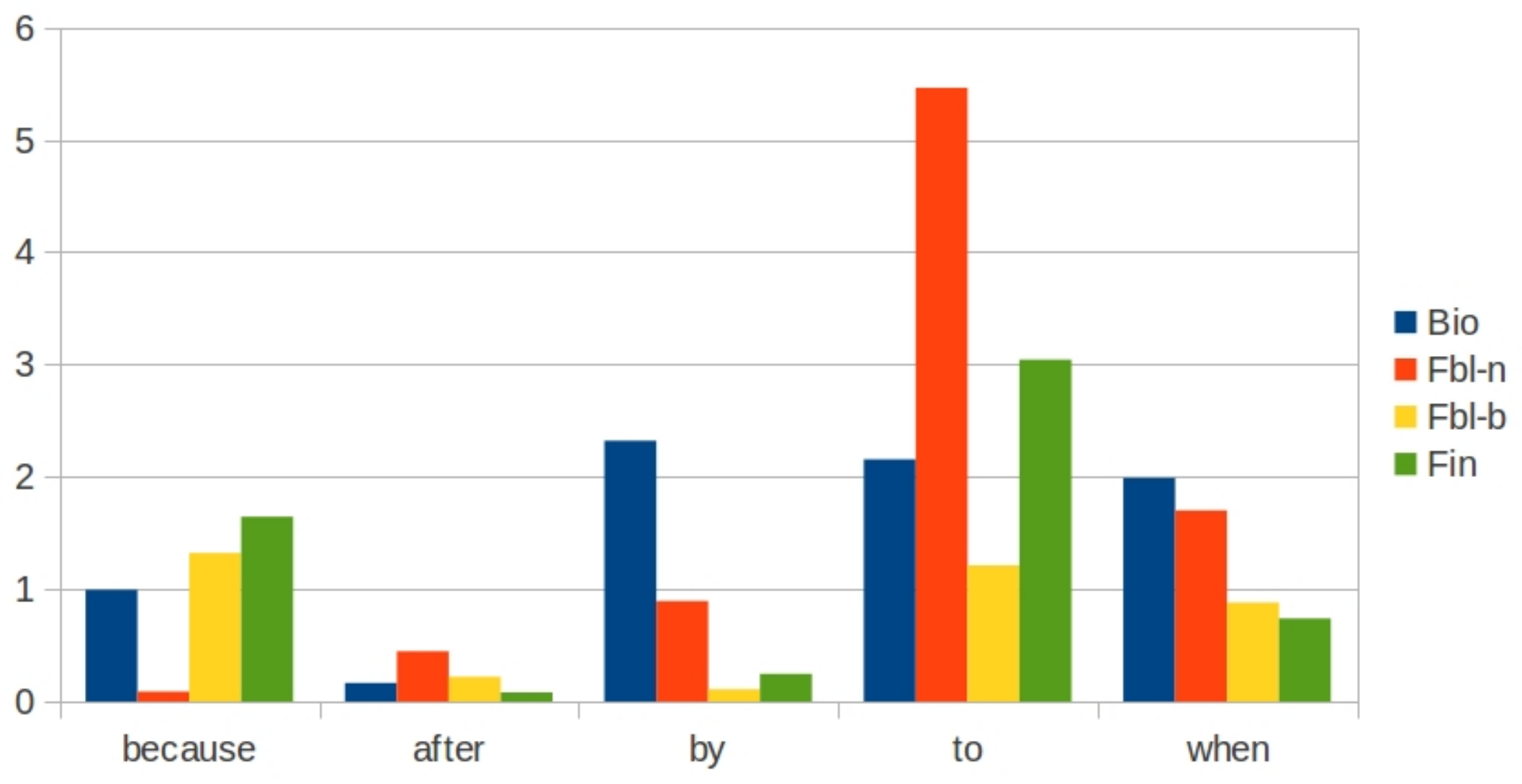
Table: Common causal markers in *Fbl-n* and *Bio*

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,

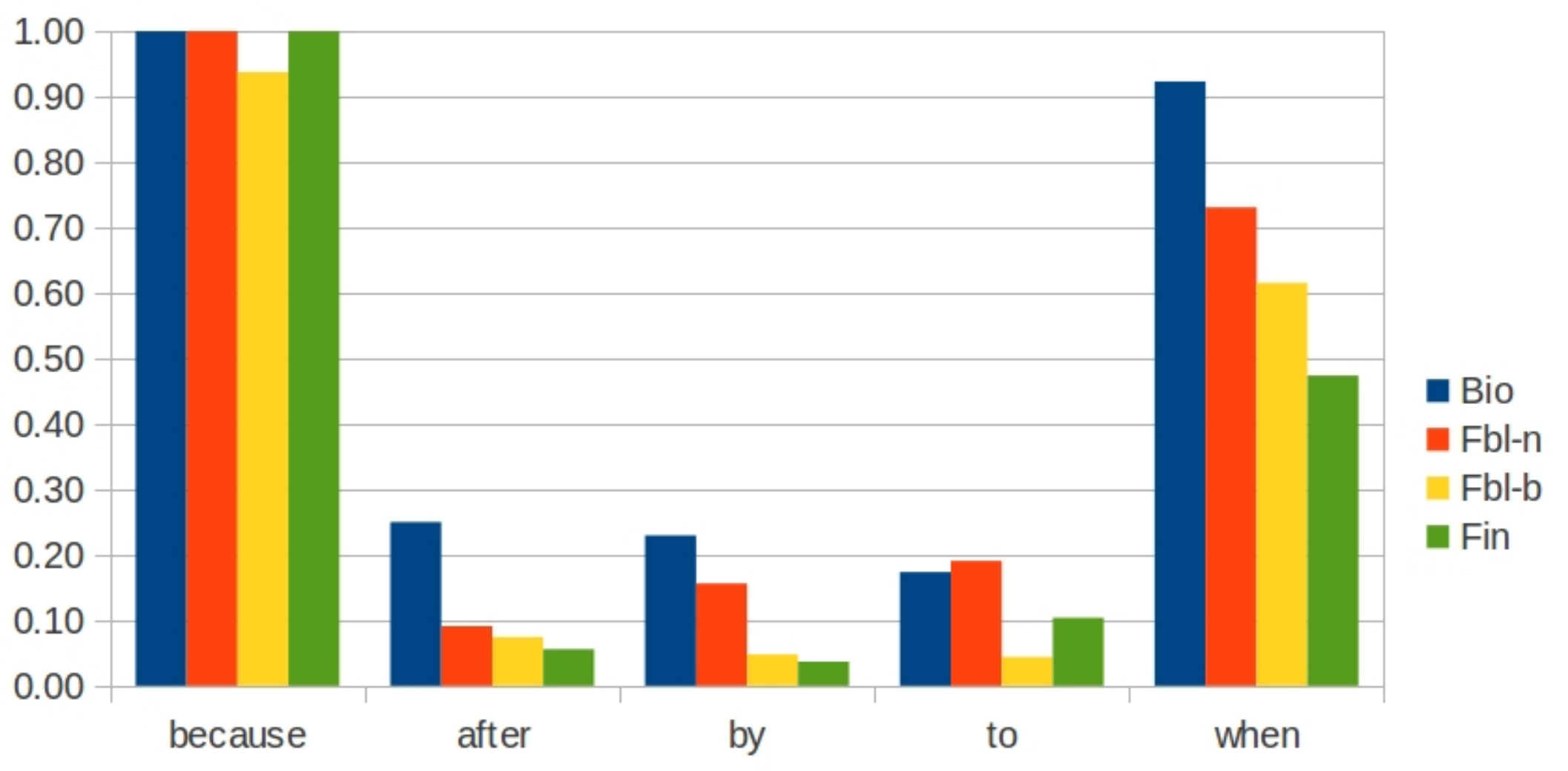
Table: Common causal markers in *Fin* and *Fbl-n*

Experiment	Fbl-n	Fin	Bio	Fbl-b
Fbl-n	-	22%	12%	22%
Fin	22%	-		
Bio	11%		-	
Fbl-b	56%			-

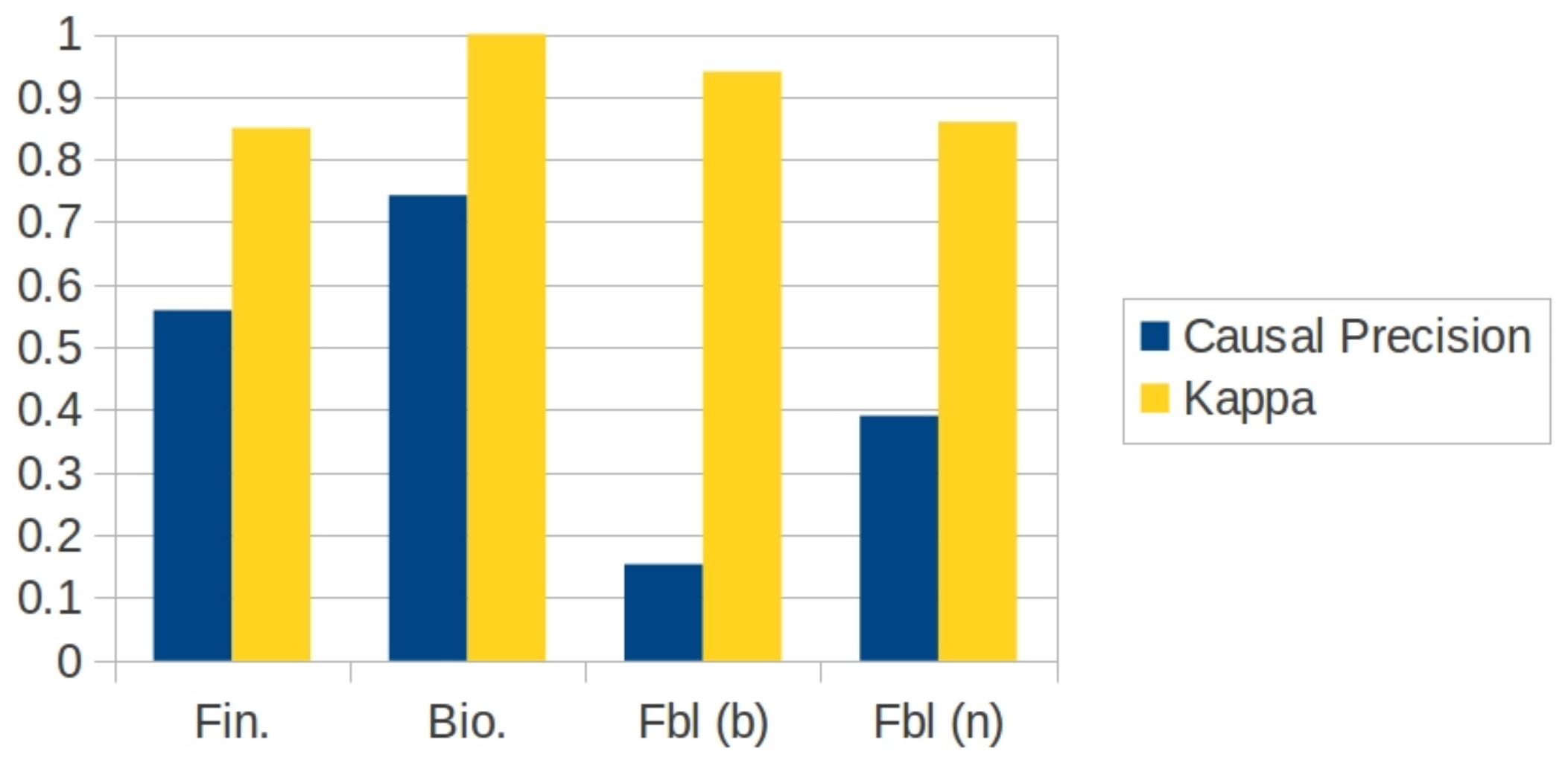
Table: Term Similarity for All Domains



Normalized Causal Frequency per Domain for Common Causal Markers



Normalized Causal Precision per Domain for Common Causal Markers



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 (Mulkar-Mehta 2011) we were unable to use the domain independent causal markers used in TREC-QA evaluation task by Prager et al. (Prager 2004) for our task of causality detection, and the causal markers needed to be modeled 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.