Enabling Search and Collaborative Assembly of Causal Interactions Extracted from Multilingual and Multi-domain Free Text

[†]George C. G. Barbosa, [‡]Zechy Wong, [♀]Gus Hahn-Powell, [♀]Dane Bell, [‡]Rebecca Sharp, [‡♀]Marco A. Valenzuela-Escárcega, [‡♀]Mihai Surdeanu
[†]Centre for Data and Knowledge Integration for Health (CIDACS), Salvador, Brazil
[‡]University of Arizona, Tucson, Arizona, USA
[♀]LUM.AI, Tucson, Arizona, USA
{gcgbarbosa,zechy,bsharp,marcov,msurdeanu}@email.arizona.edu
{ghp,dane}@lum.ai

Abstract

Many of the most pressing current research problems (e.g., public health, food security, or climate change) require multi-disciplinary collaborations. In order to facilitate this process, we propose a system that incorporates multi-domain extractions of causal interactions into a single searchable knowledge graph. Our system enables users to search iteratively over direct and indirect connections in this knowledge graph, and collaboratively build causal models in real time. To enable the aggregation of causal information from multiple languages, we extend an open-domain machine reader to Portuguese. The new Portuguese reader extracts over 600 thousand causal statements from 120 thousand Portuguese publications with a precision of 62%, which demonstrates the value of mining multilingual scientific information.

1 Introduction

The number of scientific publications has increased dramatically in the past few years. For example, PubMed[1], a repository of biomedical papers, now indexes more than one million publications per year, for a total of over 29 million publications processed to date[2].

Given this vast amount of information, it is clear that search must be a key part of the scientific research process. However, we argue that search tools today do not support this process properly. We see at least three limitations. First, most search tools tend to be relatively shallow (i.e., relying on keywords or topics), while information needs in science often require *semantics*. For example, scientific hypotheses in many sciences can be represented as causal statements, e.g., "what causes malnutrition?", or "what are the effects of pollution?". Such queries are not easily supported by current tools. Second, many sciences are becoming increasingly *multilingual*, as key scientific analyses are published in non-English venues. For example, Brazil has reduced the under-5 mortality rate resulting from poverty-related causes through its *Bolsa Familia* program (BFP), a widespread conditional money transfer to poor households (Rasella et al., 2013). However, most of the data collected in the BFP and the resulting analyses are only made available through scientific reports in Portuguese. For example, SciELO[3], an electronic repository of papers published in South America, now indexes 234,596 publications in Brazilian Portuguese[4]. Lastly, research is *iterative* and *collaborative*, whereas most search is stateless and private. For example, understanding children's health requires collaborations across multiple disciplines, e.g., biology, economy, education.

We propose a system for the search of scientific literature that addresses these three limitations. In particular, the contributions of our work are:

(1) An approach for the search of causal statements that can be both direct and indirect. Our approach relies on a novel approach for open-domain information extraction (OpenIE) that is unsupervised and domain agnostic. The proposed OpenIE method relies on syntax, and performs extractions using a top-down grammar, which first extracts relevant events, followed by event arguments, whose boundaries are determined by the syntactic constraints of the event predicate. The extractions are assembled into a graph knowledge base (KB), which supports both direct and indirect searches across causal pathways.

[1] https://www.ncbi.nlm.nih.gov/PubMed

[2] As of February 4, 2019. See the Advanced search tab on the PubMed website.

[3] http://www.SciELO.br

[4] As of February 4, 2019. See SciELO analytics: https://analytics.SciELO.org

Proceedings of NAACL-HLT 2019: Demonstrations, pages 12–17
Minneapolis, Minnesota, June 2 - June 7, 2019.

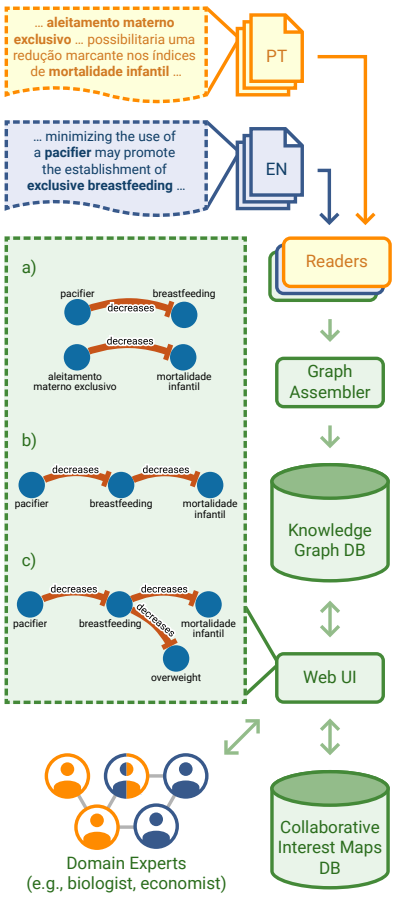

Figure 1: System architecture and example. Causal relations from sentences about breastfeeding in English (Maastrup et al., 2014) and Portuguese (Cavalcanti et al., 2015) are extracted and used by domain experts to collaboratively build a shared causal model of the task of interest, called an interest map, through a web UI. (a) A user searches for causes of *breastfeeding* and effects of *aleitamento materno exclusivo* (exclusive breastfeeding), and adds two interesting links to a shared interest map. (b) A second user merges *aleitamento materno exclusivo* and *breastfeeding*. (c) A third user adds an additional link (to *overweight*) from a new search to the shared interest map. The constructed interest maps are stored in a separate database, where they can be edited in real-time by collaborators.

(2) A multilingual search platform. We provide OpenIE grammars for English and Portuguese, and demonstrate their utility in searching PubMed and SciELO.

(3) A framework for collaborative model building. The proposed system allows end users to save the results of their semantic searches into an editable graph knowledge base, which can be shared and edited in real time by multiple collaborators. The underlying functionality for this collaborative component relies on Operational Transformations (OT), which is a conflict-free and non-blocking change propagation algorithm that allows individual users to edit a shared knowledge base in real time (Sun and Ellis, 1998).

2 Architecture

Our approach for information aggregation combines the output of machine readers into a knowledge graph which can be efficiently queried, stored, filtered, and edited by multiple users in real-time.

Specifically, given a collection of documents, we first extract relevant relations using a set of rule-based machine readers. This approach can use the output of any reader (e.g., the biomedical relation extraction framework of Valenzuela-Escárcega et al. (2018) or the open-domain framework of Hahn-Powell et al. (2017)), but here we focus on the Eidos reader (Section 3) which we extend to Portuguese (Section 3.2) in order to increase the coverage of the knowledge graph by including scientific publications in Portuguese.

The extracted concepts are unified using the deduplication approach of Hahn-Powell et al. (2017), which uses an $\mathcal{O}(n)$ hash-based approach to fingerprint relation and concept attributes paired with a set of normalized terms filtered against a series of linguistic constraints. A graph database is then populated with the unified concepts and the relations linking them. During this process, we preserve all evidence for the extractions, along with information about whether relations were hedged or negated. We employ two Lucene[5] indexes, one for indexing the content of the papers (for use in filtering the knowledge graph based on a specified context), and another for the concepts in the knowledge base (to allow for faster querying).

An important component of this system is a web-based user interface (UI) which allows users to query the graph easily and incrementally select results in order to construct a qualitative influence model, which we refer to as an *interest map* (Section 4). This UI features a real-time collaborative graph editor that is conflict-free and non-blocking, allowing multiple users to work together on a shared interest map.

3 Reader

In the context of OpenIE (Banko et al., 2007), determining the fixed set of relevant entities and events, and aggregating this information across domains and languages is likely impossible. For this reason, we use the Eidos reader (Sharp et al.,

[5] https://lucene.apache.org

2019), which is a taxonomy-free OpenIE system that uses a top-down information extraction pipeline. This pipeline begins by finding relations of interest such as causal statements (through the use of specific trigger words), and continues by extracting the concepts that participate in these relations from the syntactic context.

3.1 Reading with Eidos

To understand the individual steps of Eidos's top-down approach, consider the example sentence, *According to two studies, breast milk with omega-3 LCPUFA reduced allergic manifestations.*

First, the Eidos system finds causal and correlation relations, using a set of trigger words with a grammar of rules written in the Odin information extraction framework (Valenzuela-Escárcega et al., 2016). Odin consists of a declarative language, capable of describing patterns over surface and syntax, coupled with a runtime engine that applies these rules in a cascade, making the previous matches available for subsequent rules. In the sentence above, a Causal relation would be triggered by the predicate *reduced*, with an initial cause of *milk* and an initial effect of *allergic manifestations*. Using the approach of Hahn-Powell et al. (2017), Eidos then expands these initial arguments by traversing outgoing dependency links (with some exceptions such as conjunctions). For example, here *milk* is expanded to *breast milk with omega-3 LCPUFA*. The final system output of the Eidos system for the sentence above is shown in Figure 2.

3.2 Extension to Portuguese

We adapted the English-based Eidos system to extract causal relations from Portuguese text by first translating the trigger words and words related to filtering, negation, and hedging. We compared the syntactic preprocessing of a sample of causal sentences in English with their Portuguese translations, writing additional rules to account for differences. Rules were also written to capture lexicalized causal patterns in Portuguese. During rule development, we ran the reader over a 1K article sample of SciELO multiple times, evaluating the accuracy of each rule and adjusting them to remove incorrect extractions.

Since the grammars that Eidos uses operate over universal dependency (UD) syntax (Nivre et al., 2016), and are largely unlexicalized (with the exception of certain prominent causal forms, e.g., *due to*), we anticipated that minimal adjustments to the grammars would be needed. However, the Portuguese UD dataset used v2 of UD, while the grammar for English was written for UD v1[6]. Thus, some relations were tagged differently between the two languages, for example, `nmod` relations for English were split into `nmod` and `obl` in Portuguese. Because the Portuguese training data for UD was considerably smaller than for English, we also had to deal with the lower accuracy of the dependency parser[7], which represented a challenge when porting the grammars.

In total, we ported eight high-yield rules to Portuguese. An analysis of the extractions from the 1K article sample showed that approximately 65% of the extractions were made by a single active voice rule whose arguments are matched by traversing `nsubj` and `obj` dependencies. The next most frequently used rule, which matches causal events where the trigger is followed by the token *por*, e.g., *diminuído por* [reduced by], accounted for 15% of the extractions. No other rule accounted for more than 5% of the extractions.

Note that the Portuguese extractions are currently kept separate from the English ones. That is, the user must explicitly search for causal pathways by language. However, these results may be manually aggregated in the collaborative model workspace, described below. We describe possible strategies for the automated integration of cross-language results in Section 6.

4 Collaborative model builder

The causal graphs built from the extractions are useful for finding direct and indirect relations between concepts. However, in order to truly support the scientific research process, we argue that the resulting system must implement the following additional functionality:

(1) It must support *iterative* and *stateful* search. It is unlikely that any single search query solves a real-world research problem. It is thus necessary to allow multiple searches whose outputs are saved in the same state or model. For example, Jensen et al. (2017) showed that understanding children's health requires information from biology, psychology, economy, and environmental science.

[6] https://universaldependencies.org/

[7] The parser used for this component was an ensemble of Malt parsers (http://www.maltparser.org), as introduced in Surdeanu and Manning (2010).

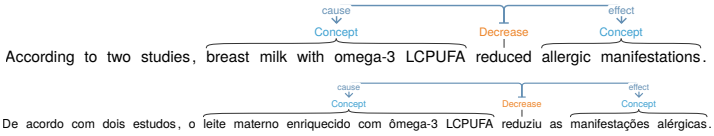

Figure 2: TAG (Forbes et al., 2018) visualization of the multilingual reader's output for one sentence in English and Portuguese.

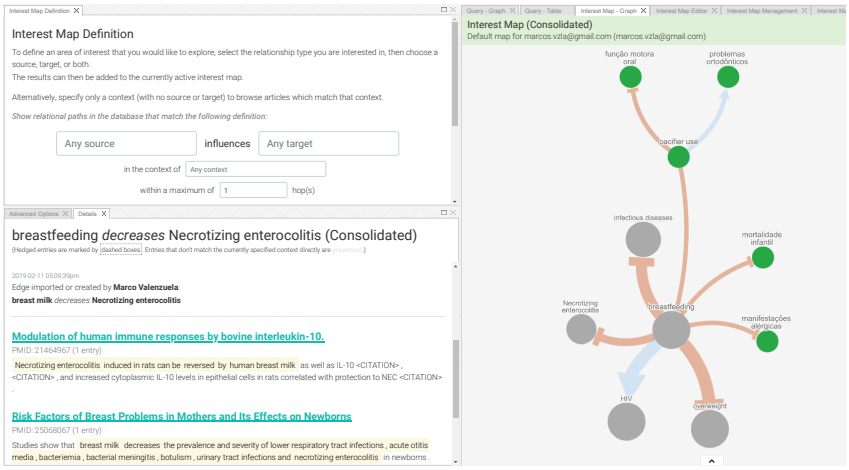

Figure 3: Screenshot showing some key functionality of our system's user interface. In the upper left panel, users can search for direct or indirect causal statements found in the literature. The right panel has the multilingual interest map, collaboratively built by different users from multiple search results. The bottom left panel has the evidence for a selected causal interaction.

(2) It must allow the addition of *background knowledge* that is known to the domain experts, but is not published in literature.

(3) Most importantly, the above operations must be performed in a *collaborative* environment that allows multiple experts to contribute to the same model, or interest map, in real time. The National Science Foundation has recognized that inter-disciplinary collaborations have become a fundamental aspect of science and has called for "growing convergence research" in its "10 big ideas".[8]

To implement the above functionality, we added a module for collaborative model building, which incorporates: (a) the ability to incrementally save the results of causal searches in a user's *interest map*, thus accumulating (a subset of) search results that capture the problem of interest; (b) operations to edit this interest map such as adding causal relations (to account for the user's background knowledge) and deleting them (to account for machine errors); and (c) real-time collaborative functionality, which allows users to share their interest maps, and edit them in parallel, in real time.

The real-time collaborative functionality is implemented using Operational Transformations (OT), which is a conflict-free and non-blocking change propagation algorithm that allows individual users be able to edit without waiting on others even under high-latency (Sun and Ellis, 1998). Typically OT is applied to documents (e.g., as with Google Docs), but here we apply it to our interest maps that are represented as directed causal graphs.

Briefly, each client has a local copy of the shared interest map, which they are free to edit. The edits are represented as *operations* (e.g., deletion of a node, or addition of a relation link). Operations generated by different clients are each transformed according to the operations of the other clients in order to synchronize the interest maps. The result is an intuitively-built unified interest map that incorporates the input from all expert users, without requiring them to be concerned with manual synchronization or conflict resolution.

5 Discussion

As shown in Figure 3, the collaborative model builder summarized in the previous section enables users to aggregate influence statements from multiple searches, multiple domains, and multiple languages. This allows end users to make full use of any information complementarity (i.e., between different domains or different languages) that is inherent in inter-disciplinary research.

Table 1 shows overall statistics for the two document collections currently processed. The table indicates that both collections contain approx-

[8] https://www.nsf.gov/news/special_reports/big_ideas/

	English	*Portuguese*
Documents	94,684	121,801
Concepts in causal interactions	1,550,912	772,470
Causal interactions	2,121,574	631,965
Precision	54%	62%

Table 1: Statistics of English and Portuguese document collections, including number of causal interactions, and number of concepts participating in such interactions. Precision was computed over a sample of 50 statements in each language. We considered an interaction to be correct if the sentence supports the interaction, the polarity (promotes/inhibits) and direction of the interaction are both correct, and the spans of the two arguments overlap with the correct spans.

imately 100K documents (more for Portuguese, less for English), and the readers extracted 2.1M causal statements from the English documents with a precision of 54%, and 631K causal statements in Portuguese with a precision of 62%, which demonstrates the value of mining multilingual scientific information. In this evaluation, extracted causal relation arguments were considered correct if the argument extracted overlapped with the correct argument. For example, in the sentence "IL-10 decreases epsilon transcript expression," the strictly correct extraction would be: (IL-10; decreases; epsilon transcript expression). Based on our evaluation criteria, the following extraction would be also considered correct (IL-10; decreases; epsilon transcript), as the span of the second argument overlaps with the strictly correct argument.

The difference in precision between Portuguese and English might be due to the fact that the Portuguese reader uses a smaller set rules that extracts approximately 4 times fewer causal statements than the English reader. Additionally, the evaluation was performed on a sample of 50 extractions, and so the difference may not be statisticaly significant.

6 Future work

Our future work efforts will focus on extending the multi-linguality of the proposed system. Given the architecture currently in place, we predict that extending it to other languages will not be too costly. We plan to use the corpora from the Universal Dependencies effort[9] to train part-of-speech taggers and syntactic parsers for additional languages. Our semantic causal grammars are mostly unlexicalized; most of the effort required to adapt them to other languages will be on translating the causal triggers.

In order to merge knowledge graphs constructed using corpora from different languages, we will need to align of multilingual terminology. Some domains may already provide manually translated vocabularies, for example, within the medical domain, UMLS (Bodenreider, 2004) provides translations of the controlled vocabulary MeSH (Lipscomb, 2000) for several languages. For domains in which manual translations are not available, we can take advantage of recent developments in unsupervised bilingual dictionary induction (Conneau et al., 2017; Kementchedjhieva et al., 2018) to learn alignments.

Lastly, we will work on methods to minimize the spreading of accidental misinformation, which may be introduced by incorrect extraction or statements that are not factual. To mitigate the former issue, we found that extraction redundancy provides a strong signal, i.e., statements extracted multiple times from different publications are more likely to be correct. For the latter, we will employ recently-proposed methods for factuality detection (Rudinger et al., 2018).

7 Conclusion

We introduced a novel system[10] that facilitates the search for multilingual and multi-domain causal interactions that are either direct or indirect. Further, the proposed system includes a framework for collaborative model building, which allows multiple domain experts to collaborate in real time on the construction of a causal model for a given problem, which aggregates the results of multiple searches as well as background knowledge manually added by the experts.

Acknowledgments

This work was funded by the Bill and Melinda Gates Foundation HBGDki Initiative. Marco Valenzuela-Escárcega and Mihai Surdeanu declare a financial interest in LUM.AI. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

[9] https://universaldependencies.org

[10] https://multiling.demos.clulab.org

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Sandra Hipólito Cavalcanti, Maria de Fátima Costa Caminha, José Natal Figueiroa, Vilneide Maria Santos Braga Diegues Serva, Rachel de Sá Barreto Luna Cruz, Pedro Israel Cabral de Lira, Malaquias Batista Filho, et al. 2015. Fatores associados à prática do aleitamento materno exclusivo por pelo menos seis meses no estado de pernambuco. *Revista Brasileira de Epidemiologia*, 18:208–219.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Angus Graeme Forbes, Kristine Lee, Gus Hahn-Powell, Marco Antonio Valenzuela-Escárcega, and Mihai Surdeanu. 2018. Text annotation graphs: Annotating complex natural language phenomena. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resources Association (ELRA).

Gus Hahn-Powell, Marco A. Valenzuela-Escárcega, and Mihai Surdeanu. 2017. Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. *Proceedings of ACL 2017, System Demonstrations*, pages 103–108.

Sarah KG Jensen, Anne E Berens, and Charles A Nelson 3rd. 2017. Effects of poverty on interacting biological systems underlying child development. *The Lancet Child & Adolescent Health*, 1(3):225–239.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. In *CoNLL*.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Ragnhild Maastrup, Bo Moelholm Hansen, Hanne Kronborg, Susanne Norby Bojesen, Karin Hallum, Annemi Frandsen, Anne Kyhnaeb, Inge Svarer, and Inger Hallström. 2014. Factors associated with exclusive breastfeeding of preterm infants. results from a prospective national cohort study. *PloS one*, 9(2):e89077.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Davide Rasella, Rosana Aquino, Carlos A. T. Santos, Romulo Paes-Sousa, and Mauricio L. Barreto. 2013. Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of brazilian municipalities. *The Lancet*, 382:57–64.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 731–744.

Rebecca Sharp, Adarsh Pyarelal, Benjamin M. Gyori, Keith Alcock, Egoitz Laparra, Marco A. Valenzuela-Escárcega, Ajay Nagesh, Vikas Yadav, John A. Bachman, Zheng Tang, Heather Lent, Fan Luo, Mithun Paul, Steven Bethard, Kobus Barnard, Clayton Morrison, and Mihai Surdeanu. 2019. Eidos, INDRA, & Delphi: From free text to executable causal models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics.

Chengzheng Sun and Clarence Ellis. 1998. Operational transformation in real-time group editors: Issues, algorithms, and achievements. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*, CSCW '98, pages 59–68, New York, NY, USA. ACM.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference (NAACL-2010)*, Los Angeles, CA.

Marco A. Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T. Morrison. 2018. Large-scale automated machine reading discovers new cancer driving mechanisms. *Database: The Journal of Biological Databases and Curation*.

Marco A. Valenzuela-Escárcega, Gustave Hahn-Powell, and Mihai Surdeanu. 2016. Odin's runes: A rule language for information extraction. In *LREC*.