

Discovering Causal Sentences with Automatically Learned Patterns

Shreekanth Karvaje¹, Bharat Ravisekar¹, Saurav Sahay¹, Baoli Li¹,
Ernest Garcia², Ashwin Ram¹

¹ College of Computing, Georgia Institute of Technology
{shreek, rbharat, baoli, ashwin}@cc.gatech.edu

² Department of Radiology, Emory University
Atlanta, GA 30322, USA
ernest.garcia@emoryhealthcare.org

Abstract. We propose a semi-supervised method to extract rule sentences from medical abstracts. Medical rules are sentences that give interesting and non-trivial relationship between medical entities. Mining such medical rules is important since the rules thus extracted can be used as inputs to an expert system or in many more other ways. The technique we suggest is based on paraphrasing a set of seed sentences and populating a pattern dictionary of paraphrases of rules. We match the patterns against the new abstract and rank the sentences.

Introduction

Huge amounts of medical rules are scattered in unstructured format in medical publications, journal articles and web resources. Manually extracting these rules is often a labor-intensive and a time-consuming endeavor. In this paper, we propose a very flexible semi-supervised method based on bootstrapping, to extract such knowledge rules. These extracted rules can be fed to an expert system like the PERFEX [1].

We consider medical knowledge as being present in sentences containing a notion of causality. In English, the causative constructions can be explicit, semi-explicit or implicit. Explicit causative constructions contain unambiguous and relevant keywords such as cause, effect, consequence that define the type of relation. The concerned entities are further linked either using causal verbs or causative links (complex clauses and phrases). In this paper we are considering explicit and semi-explicit causative verbs as to be representing causality.

The underlying assumption we make is that if a sentence is known to be containing a useful rule, then sentences that are its paraphrases might contain some rules as well.

Method

Our method of extracting knowledge sentences is as follows:

Building triples database for paraphrasing

A large corpus (like OHSUMED) is used to build a “triples” table which is built according to the method suggested in DIRT [2]. The table contains triplets of path and slots (SlotX and SlotY); and their associated fillers, frequencies and mutual information. We use the Minipar [3] dependency parser to build a dependency tree of each sentence in the corpus and extract paths from this tree using the following criteria:

1. The path should join only content words (Nouns, Verbs, Adjectives, and Adverbs).
2. The paths start and end with entities, in the form of Nouns. This ensures that the resulting patterns link only entities.

To calculate the mutual information between a path p and its filler w , we use the Equation 1, where Slot is either SlotX or SlotY.

$$mi(p, Slot, w) = \log \frac{P(p, Slot, w)}{P(Slot)P(p | Slot)P(w | Slot)} \quad (1)$$

Building the pattern dictionary

A small set of seed sentences which represent the knowledge in medical abstracts is input to the system in this step. These sentences are used to build a pattern dictionary. The quality of the patterns extracted depends on the seed rules.

Paths (base patterns) are extracted from the seed sentences as done in the previous step. For each path thus extracted, a matching path is selected from the Triples database, provided one exists. If there is no

direct match, we expand the base pattern based on the nouns that are ending it on either side.

For each base pattern, we find a set of paraphrases using the triples table. These paraphrases form the candidate patterns. The paraphrases are extracted using a similarity score based on the similarity of slots. To calculate the similarity between any two slots $slot_1 = (p_1, s)$ and $slot_2 = (p_2, s)$ belonging to two paths p_1 and p_2 , we use the following metric:

$$Sim(slot_1, slot_2) = \frac{\sum_{w \in T(p_1, s) \cap T(p_2, s)} mi(p_1, s, w) + mi(p_2, s, w)}{\sum_{w \in T(p_1, s)} mi(p_1, s, w) + \sum_{w \in T(p_2, s)} mi(p_2, s, w)} \quad (2)$$

To calculate the similarity between paths, we propose Equation 3.

$$Similarity(p_1, p_2) = Max\left(\frac{1}{2}(Sim(p_1.SlotX, p_2.SlotX) + Sim(p_1.SlotY, p_2.SlotY)), \frac{1}{2}(Sim(p_1.SlotX, p_2.SlotX) + Sim(p_1.SlotY, p_2.SlotY))\right) \quad (3)$$

Only those candidate patterns that clear a certain thresholds ψ , θ (set empirically) for mutual information and similarity respectively are considered.

Selecting and Ranking candidate sentences in new abstract

The algorithm for this step is outlined below:

1. For each sentence in a new document, generate the dependency tree using Minipar and extract paths as done earlier.
2. Each such path extracted, use the triples database to expand this path (by using the similarity, mutual information of the slots and picking the most similar paths).
3. Assign usefulness score to the input sentence using the number of patterns that match the expanded paths. All the sentences in the document that cross a certain threshold θ are output.

Evaluations

We evaluated our proposed method on the OHSUMED corpus. OHSUMED collection contains a set of 348,566 abstracts from

MEDLINE. We obtained 127,454 unique paths, 199,844 unique *SlotX* fillers and 253,541 unique *SlotY* fillers from the training set.

A set of 38 patterns was used to seed the pattern dictionary. A list of 4,657 paraphrased patterns was extracted. We judged the quality of the patterns by manually evaluating the quality of top 100 patterns. A pattern was adjudged useful if majority of the evaluators have marked it as useful. From this experiment it was clear that on average 75% patterns extracted in the dictionary represent useful patterns.

From the test set, 300 abstracts were manually annotated for rules. Table 1 gives the precision and recall for varying values of threshold θ .

Table 1. Performance results with varying θ

θ	P	R
0	13.86	90.41
0.01	13.97	78.08
0.02	12.84	57.53
0.05	14.78	46.57

Conclusion and Future Work

We have shown that we can extract medical rule sentences with a semi-supervised technique. Our results are quite promising and uphold our assumption that if a sentence is known to be containing a useful rule, then sentences that are its paraphrases might contain some rules as well. In future, we plan to improve the system by tuning the ranking functions used at various levels in our system. Ontologies like WordNet, UMLS could also be incorporated for better accuracy.

References

1. Ezquerra N., Mullick R., Cooke D., Krawczynska E., and Garcia E. 1993. PERFEX: An Expert System for Interpreting PerfusionImages. *Expert Systems with Applications*, Vol. 6, pp. 459-468.
2. Lin D and Pantel P. 2001. DIRT—Discovery of inference rules from text. In: *Proc of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, August 26–29, 2001. ACM Press; pp 323–328.
3. Dekang Lin. 1999. MINIPAR: A Minimalist Parser. In *Maryland Linguistics Colloquium*, University of Maryland, College Park.