Lexical Resources for the Identification of Causative Relations in Portuguese Texts

Brett Drury1, Paula C.F. Cardoso1, Janie M. Thomas2, Alneu de Andrade Lopes1

1 Institute of Mathematics and Computer Sciences (ICMC), University of São Paulo (USP)
2 Federal University of São Carlos (UFSCar)

Abstract. The identification of causal relations from text is a mature problem in Natural Language Processing. There are a number of resources and tools to aid causative relation extraction in English, but there seems to be a limited number of resources for Portuguese. This paper presents a number of resources which are designed to aid the researcher and the practitioner to extract causative relations from Portuguese texts.

Key words: Lexical Resources, Causative Relations, Dictionaries, Annotated Data

1 Introduction

The extraction of causative relations is a mature task in the field of information extraction. Altenberg [1] defines a causal relation as the relation which exists between two events if one event is the cause of the other. The identification of the effect event and its cause event may allow the construction of causative models where effects can be inferred from the existence of a cause event. For example, the casual relation, A produção de papel (cause event) causa grandes danos a meio ambiente (effect event) indicates that paper production causes damage to the environment. Therefore, if a text states that there is paper production in a specific area, we may infer from that information that there will be environmental damage in that area at some future date. It is arguable that causative relation identification is not only an interesting academic exercise, but has a commercial application.

This paper will present a number of lexical resources which may aid the researcher to extract causative relations from Portuguese texts. The resources are not a comprehensive set which cover all types of causation in text, but a subset of the most commonly studied forms of causative relations. The lexical resources described are: (1) a manually constructed list of causative verbs, (2) a manually verified gold standard of causative and non-causative relations, (3) an automatically constructed list of Portuguese causative verbs, and (4) a list of automatically extracted causative relations.
The format of this paper will be the following: a brief discussion of causation, related work, a description of the manually constructed lexical resources, and a summary of the automatically constructed lexical resources.

2 Causation and Related Work

This section will briefly discuss causation in text and related lexical resources which contain causative information.

2.1 Causation

As discussed earlier, causation in text may be seen as two events sequential in time where one is the cause of the other. These two events are typically linked explicitly or implicitly in the text. There are a number of methods of linking cause and effects events in the text. They are too numerous to describe here, but we will discuss the causation types addressed by the lexical resources described in this paper.

A common form of causation linkage is to use "causative verbs", for example, "Fumar causa cancer" (Smoking causes cancer), which connects Fumar and cancer with the transitive causative verb, "causar". [21] defines a transitive causal verb (V) as "for x to V y is for x to cause some event in the history of y, or for x to cause some state of affairs which consists in y's being in some state or other". The author provides an example of a transitive casual verb which obeys this definition as x moves y. The author also states that causative verbs in addition to being transitive, may be: (1) intransitive or (2) uni-transitive. Levin[12] stated that the transitive form of a causative verb may be generalized with the following pattern: NP V NP, where NP = Noun Phrase and V = Verb. [7] categorized causative verbs as: simple, instrumental and resultative. The simple causative verb provides the link between the cause and effect events. The instrumental causative verb contains all or part of the cause event as well as the causative link. The resultative causative, in turn, includes some or all of the effect event as well as the causative link. Resultative verbs may have a sentiment category.

Another form of common linkage is "adverbial". Khoo et al. [9] defined an adverbial link as: "an adverbial which provides a cohesive link between two clauses". The authors split the adverbial link into two sub-types: (1) anaphoric adverbial and (2) cataphoric adverbial. According to the authors, an anaphoric link has an anaphoric reference to the preceding clause and a cataphoric adverbial link has a cataphoric reference to the following clause.

A detailed discussion of causation may be found in [1] and [3].

2.2 Related Work

The literature search for related work was focused on lexical resources which could be used for causative relation extraction. Lexical resources are pre-compiled
data sources which have had annotations, categorisations, relations, etc., added
to it.

Certain Levin verb classes[12] may have members which are causative. Verb-
Net[19] is a physical representation of Levin classes with additional refinements
and sub-classes. It also has mappings to other lexical resources such as Word-
Net[14]. WordNet[14] is a taxonomy of words constructed from synsets. It con-
tains a cause relation which links: verb causative pairs such as fell-fall and point-
ers from causative verbs to the corresponding anti-causative intransitive sense of
the same word. VerbOcean[4] is a lexical resource to classify verbs into semantic
groups. The groups are: (1) similarity, (2) strength, (3) antonymy, (4) enable-
ment, and (5) happens-before. A full description of each class is provided by
[4]. The enablement and happens-before classes are synonymous with the cause
relation in WordNet.

The lexical resources for identifying verbs with causative properties has thus
far been limited to English. However, there are a number of lexical resources
which have been designed for non-English languages, for instance, Papel[17]
which is a resource for Portuguese and has causative relations for specific verbs.
AnCora[2] is a lexical resource similar to VerbNet for the Spanish and Catalan
languages. The resource is constructed from predicates which are related to one
or more semantic classes. The relations are determined by the predicates senses.
The classes are differentiated by four event classes: accomplishments, achieve-
ments, states and activities.

3 Manually Created Lexical Resources

This section will discuss the creation of manually created resources for causative
relation extraction. The resources presented in this section are: 1. A gold stan-
dard of causative relations and non-causative relations and 2. A list of simple and
resultative causative verbs. The resultative causative verbs have been classified
into sentiment categories.

3.1 Gold Standard

A gold standard, in this instance, is a small number of causative relations which
have been annotated by six annotators. The data set used for the annotation had
a 1,000 sentences which were randomly selected from a larger corpus of 300,000
sentences. The annotation guidelines were classify the sentence as either: (1)
causative or (2) non-causative. If the sentence contained causative information,
the annotators were to (1) identify a cause and an effect event and (2) identify a
causative link between the cause and effect events. This linkage could be a single
word or a multi-word expression. If, however, the causative link was implicit
then this annotation could be blank. The cause and effect events could be either
in: the same sentence or separate sentences.
The gold standard of causative relations were annotated with multiple annotators using a web-based annotation tool \(^3\) [6]. The annotators ranged from experienced annotators to novices.

**Lexical resource** consists of two types of annotations: sentence classification and sentence annotation, organized in separate files. Each file contains one sentence per line. The sentence classification file contains a sentence and a manually added category of either: "non-causative" or "causative". The sentence annotation file contains a sentence which has the following annotations: cause, causative link and effect.

**Gold standard characteristics** for this gold standard are measured by the number of annotated sentences and their quality. The quality of sentence category classification was measured by: the percentage annotator agreement (PAA) and Cohen kappa co-efficient[20]. PAA may be represented by: \(\frac{naSC}{nSC}\) where \(naSC\) is the number of annotators who agree on the majority classification for a specific sentence and \(nSC\) is the total number of sentence classification. The weakness of this approach is that annotators could agree on a classification by chance, i.e., on a two-class classification problem, the chance of two annotators who are classifying sentences randomly agreeing on the same category is 0.5. The Cohen kappa co-efficient attempts to mitigate chance agreement and is represented as:

\[
\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},
\]

where \(Pr(a)\) is the observed relative agreement between annotators and \(Pr(e)\) is the hypothetical probability of chance agreement between annotators. The sentence classification task yielded 833 sentences which had a PAA of 77.96% and a Cohen Kappa Co-efficient of 60.74.

The agreement of the annotators identification of: cause, effect and causative link were measured by an average Levenshtein distance[15](ADL). A Levenshtein distance may be used to calculate the similarity between two texts. An ADL was calculated by estimating a Levenshtein distance for each pair of annotations made by the different annotators on the same sentence. An average is taken for all of the calculated Levenshtein distances. An ADL was calculated for cause, effect and causative link. The higher the ADL value, the higher the similarity of the annotations. The ADL for the 379 annotations were: 0.73 (±0.32) for cause events, 0.68 (±0.29) for effect events and 0.71 (±0.35) for the causative link. The values demonstrate that the annotators were in broad agreement, but there were minor differences in the annotations.

### 3.2 Lists of causative verbs

As discussed before, a form of causation may be represented using verbs and they make take the form of: simple, resultative and instrumental. Resultative may have a sentiment category. The lexical resource contains three lists of causative

---

\(^3\) The annotation platform is accessible at [http://goo.gl/d2UN93](http://goo.gl/d2UN93).
verbs: simple, resultative (negative) and resultative (positive). We did not find any instrumental causative verbs.

The construction of the verb lists initially selected verbs from the annotated data previously described. The verbs were split into: simple, resultative (negative) and resultative (positive) categories. The lists were expanded using synonyms from Onto.pt [8]. The verb candidates were evaluated by two annotators who labelled them with a category: simple, resultative (negative), resultative (positive) or not causative verb (NCV). The verb candidates which were labelled by at least one annotator as NCV were eliminated. Verbs which had full annotator agreement were added to the appropriate verb list.

4 Automatically Constructed Lexical Resources

The weakness of manually constructed resources is that it is a slow process which realizes a small number of resources. An alternative is to use an automatic method which relies upon a small amount of labelled data. In this section, we will describe a process which labels causative relations and categorizes verbs into non-causative and causative categories.

The described process identifies explicit causation, i.e., the causative link between the cause and effect events is explicitly stated. In addition, we identified a subset of causation which uses causative verbs to provide the link between the cause and effect events.

4.1 Overview of Algorithm

The rationale of the algorithm is to detect causative relations which obeys Levin's[12] pattern of $NP \quad V \quad NP$, where $NP$ = Noun Phrase and $V$ = Verb, for transitive causative verbs. The algorithm classifies the verbs and a series of decision rules label the NPs.

As stated earlier, the amount of manually labelled data was limited; consequently, a classifier which is induced from this data will be weak and may erroneously label verbs. We, therefore, decided to use a semi-supervised learning technique called self-training. Self-training uses a base classifier induced from labelled data to label unlabelled data. Unlabelled data which has a high confidence classification is added to the training data which will be used to induce a classifier in the next iteration. Self-training may impair a classification process by propagating errors, consequently, we used two classifiers which have two separate "views" of the data. These two classifiers need to agree on a classification before an unlabelled instance is used as a training example. This type of algorithm is known as multi-view learning[18].

We chose two classifiers: relative link classifier (RLC)[5] and a conditional random field. The RLC is a graph-based approach which extracts verbs and their arguments from labelled and unlabelled data. Each word is represented on the graph as a node. A verb is joined to its arguments by a vertex. A verb extracted from the labelled data will have a label of either causative or non-causative.
The verb arguments are removed and verbs which have common arguments are joined by a vertex. The labels from the labelled verbs are passed to the unlabelled verbs with a simple propagation strategy. A full description of this technique is provided by [5].

We also constructed a series of decision rules which labels arguments of verbs as either: effect or cause events. These decision rules are used in conjunction with the RLC to label causative relations in a sentence.

The second classifier was a Conditional Random Field (CRF) [11]. A CRF was chosen because they have been used for causative relation extraction [13]. The CRF classifies the words into several classes: non-causative, cause, effect and causative verb. The features for the CRF were chosen using a genetic algorithm feature selection strategy [16].

4.2 Algorithm Description

As described earlier, the algorithm is a self-training which uses multi-view learning that uses a global (RLC), a local classifier (CRF) and a rule labeller. The CRFs are “stacked” [10]. Stacking shows randomly selected equal divisions of the training data to \( n \) classifiers, where \( n > 1 \). A classification is returned by the stacked CRF if the separate CRF models agree on the classifications. In this case we used 3 classifiers. This number was chosen because it was found to produce the best results.

The algorithm initializes the stacked local and global classifiers using manually labelled data. The classifiers classify words of individual sentences into: cause, effect, causative verb, non-causative categories. If the global classifier and rule labeller agree with the stacked CRFs, then the classification is accepted and is added to the training data for the next iteration. At the end of the iteration, the global and local classifiers are updated and the process continues. The algorithm terminates when there are no new training candidates. The algorithm is described in full in Algorithm 1.

4.3 Lexical Resources

The lexical resources produced by the aforementioned algorithm were: labelled instances of explicit causative relations and a list of causative verbs. The algorithm produced a list 4103 labelled instances and 106 causative verbs.

The causative relations file contains one sentence per line. Each word contains a Part of Speech (POS) tag, and on occasion a annotation label of either: CV (causative link), EN (Effect phrase) or CN (Cause phrase). The causative verb list contains one verb per line.

5 Conclusion

This paper presents a group of lexical resources which have been designed to assist the researcher to identify causative relations in Portuguese texts. These resources may be obtained from http://goo.gl/2e0csd.
Input: UL, LD, DR
Output: LD
/* UL = unlabelled data, LD = labelled data, DR = decision rules */
while True do
    gc ← train(LD)
    crf ← train(LD)
    /* gc = RLC, crf = Conditional Random Fields (stacked) */
    count ← 0
    for sentence ∈ UL do
        /* test if sentence is in labelled data */
        if sentence in LD then
            continue
        end
        /* test agreement for verbs v, cause c and effect e */
        e, c, v = classify(DR, gc, sentence)
        e,1, c,1, v,1 = classify(crf, sentence)
        if e == e,1 and c == c,1 and v == v,1 then
            count ← count + 1
            /* Add training candidate to labelled data */
            LD ← appendData(LD, e, c, v)
        end
    end
    /* Termination Condition */
    if count == 0 then
        return LD
    end
end

Algorithm 1: Self-training algorithm

5.1 Acknowledgements

This work has been supported by FAPESP, grant number: 11/20451-1.

References


