Conference on Empirical Methods in Natural Language Processing

EMNLP 2015

Author Response

Title: "A Spousal Relation Begins with a Deletion of engage and Ends with an Addition of divorce": Learning State Changing Verbs from Wikipedia Revision History

Authors: Derry Tanti Wijaya, Ndapandula Nakashole and Tom Mitchell

Instructions

The author response period has begun. The reviews for your submission are displayed on this page. If you want to respond to the points raised in the reviews, you may do so in the box provided below.

Please note: *you are not obligated to respond to the reviews*.

Review #1

Appropriateness: 5
Clarity: 3
Originality: 3
Soundness / Correctness: 4
Impact of Ideas / Results: 3
Meaningful Comparison: 3
Substance: 3
Replicability: 3
Recommendation: 3

Comments

This paper approaches the task of learning state-changing verbs in an interesting way. By learning mappings between text content differences and infobox differences on Wikipedia pages, the paper learns which verbs cause which types of state changes. While the approach is interesting, the model is comparatively simple where weights are learned for the verbs and their subjects/objects in the text content difference. A mixed-integer program is used to reduce noise. The F1 of the approach seems reasonable (around 80%) but there are no baselines, so it's hard to evaluate -- adding (at least) a majority class baseline would be informative. There are some clarity issues describing the approach. These were ultimately resolvable but could be described more simply/clearly. The paper plans to distribute its dataset which should aide future research on this task. Overall, this is a nice contribution to an interesting task but I don't think it is ready for publication in its current form.

Minor points:

Sec. 1: in 3rd paragraph, might be good to say "approximately the same time" instead of "same time"

Sec. 2.1: should use gender neutral pronouns ("his or her" or "their") for people in the set of the entities P

Sec. 2.1: The notation here is a bit verbose and could be simplified. For example, in d_{p,t_p}, the second p can be dropped -- the same goes for r and ΔS. In the 4th paragraph (describing ΔS), the notation creates confusion and it may be better to describe these processes in words. As it stands, it severely hurts readability and clarity.

Sec. 2.1: The 4th paragraph mentions verbs, subjects, and objects. Please add more details about how the syntactic analysis was performed.

Sec. 2.2: Is it the case that each w_y is a vector of length $|V|$? Please clarify.

Sec. 2.3: "other birth-related updates" -- please add an example.

Sec. 2.3: How many constraints were used?

Sec. 2.3: Was any regularization used for the MaxEnt model to reduce noise? While operating in a different way than the mixed-integer program, it would be interesting to know whether it provides any help.

Typographical:

Sec. 1: Two periods after first sentence, period in 2nd sentence should be after citations. Space before period in 3rd paragraph.

Sec. 4: should be spaces before the parentheses in a citation (e.g., "categorization(Daxenberger and Gurevych, 2013)")

Sec. 4: periods are wrong around the second block of citations

Review #2

Appropriateness: 5
Clarity: 5
Originality: 4
Soundness / Correctness: 3
Impact of Ideas / Results: 4
Meaningful Comparison: 3
Substance: 4
Replicability: 5
Recommendation: 4

Comments

This paper presents a weakly-supervised approach to an interesting task on learning state-change verbs. The states are defined via existing knowledge bases. The main novelty is to use Wikipedia editing history as the learning and evaluation resource, which is potentially rich and widely accessible. The prediction model is straightforward (multi-class logistic regression). The methodology-wise contribution is to use intra-state constraints to denoise the weak supervision signal. Empirically the proposed constraints improve precision while suffer a little bit loss of recall hence result in roughly 2% absolute improvement of F1 over the base model.

There are several aspects of this paper that I like given it is a short submission. First, time-scoping knowledge base facts is no-doubtly important. Given that most state-of-the-art work on knowledge base population still focus on predicting true or false in general, I think this paper should be one of the pioneer work on this task. Second, the application of wiki-editing history is novel and very reasonable. In this sense the work may also be seen as one that learns from community knowledge, which is noisy and hard to utilize but could benefit many IE tasks. It can open up many questions in the following work.

Meanwhile I found several Cons of the paper:

1. Conducting a plain multi-class classification over all the difference documents might be too coarse. Especially, some states once established never change, such as "birthplace, death place"; some are very likely to change over time, such as "workplace"; some states transit from one to another such as "engagement" and "marriage". It is unclear one should do predictions for all of them in a unified framework.

2. Wiki-editing history has some unique characteristics that would make domain adaptation to plain text difficult. For example, it contains deletions, which do not appear in common text resources. Besides, the deletions might be performed merely to maintain the coherence and conciseness of the KB/article and do not serve as semantically informative signals.

3. Since the authors plan to release the dataset, I would like to see more statistics and analysis about the data they prepare. Regarding this I have the following questions:

(a) 16,909 person entities is a very small amount out of five years of wikipedia articles. Hence I wonder if more filtering has been done besides those mentioned in section 2.1, say, did you filter short/bad articles?

(b). I would like to see the complete set of KB states over which the methods are evaluated on in the experiment section.

(c). I would like to see after labeling how bias the data is, that is, how many positive examples vs. negative examples? Without (b)&(c) it is difficult to judge the difficulty and usefulness of the dataset.

(d). Is the 9:1 train/test split by entity or by documents? Based on the description it seems to be by document. In my opinion by entity is a better strategy since it would reduce the overlap between train/test data.

(e). To provide authentic evaluation the authors manually went through the test data and discard documents with incorrect inboxes. I would like to confirm if this means all the 20,000+ test documents are manually examined since this is a daunting amount. After this filtering, how did the statistics change? Are the test data biased in some way by the infobox errors? Besides, I think manually examining test data SHOULD NOT be performed by the authors themselves and should go to other annotators.

Overall I like the task and main approach of the paper and I think it has much space of improvement. I would remain concerned if the authors cannot address the questions regarding data preparation since more dataset doesn't necessarily benefit the community and quality control is important.

Review #3

Appropriateness: 5
Clarity: 4
Originality: 4
Soundness / Correctness: 4
Impact of Ideas / Results: 3
Meaningful Comparison: 3
Substance: 3
Replicability: 3
Recommendation: 4

Comments

This paper looks at Wikipedia revision history and uses verbs to predict state changes in relations, e g., begin-birthplace. The proposed MaxEnt+MIP method achieves a F1 at above 80. The method is interesting in that it pairs revisions with infobox edits, and might be able to find more accurate matches comparing to previous distant-supervision-based relation extraction methods. This is partially confirmed by its relative high P/R/F1.

Would the model perform better if trained with more features (e.g., dependency paths, entity types, bag of words, etc)?

Would the learnt model, if used as a relation extraction model, apply to domains other than Wikipedia text? Wikipedia has a particular style that might make learning easier compared to other text such as news. However, being applicable in news certainly make the proposed method more interesting.

Submit Response

Use the following box to enter your response to the reviews. Please limit your comments to 600 words (longer responses will not be accepted by the system).

Submit
