"A Spouse Begins with a Deletion of *engage* and Ends with an Addition of *divorce*": Learning Verbs and State Changes for Knowledge Base Updates

Derry Tanti Wijaya
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
dwijaya@cs.cmu.edu

Ndapandula Nakashole
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
ndapa@cs.cmu.edu

Tom M. Mitchell
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
tom.mitchell@cs.cmu.edu

Abstract

Most knowledge bases (KBs) that have emerged in recent years are static. They contain facts about the world yet are seldom updated as the world changes. This paper proposes a method for learning state changes brought about by state-changing verbs on its arguments (i.e., entities). Entities' state change can be viewed as updates of KB facts pertaining to the entities. We propose to learn automatically state changes brought about by verbs on entities using Wikipedia edit histories of the entities. When a state-changing event happens to an entity, the Wikipedia infobox containing facts of the entity may be updated. At the same time, text containing verbs that express the event may also be added/deleted from the entity's Wikipedia page. In this paper we use edit histories as weakly supervised data for learning verbs that relate to infobox changes. We also leverage fact-specific constraints such as mutual exclusion and relatedness among infobox slots to effectively learn infobox changes that are brought about by the addition and/or deletion of verbs in the Wiki page. From our experiments, we observe that when state-changing verbs is being added/deleted to a person's Wikipedia text, we can update infobox facts about the person effectively (with an 89% precision and 74% recall).

1 Introduction

In recent years there has been a lot of research on extracting relational facts between entities and storing them in knowledge bases (KBs). These knowledge bases such as YAGO (which extract facts from Wikipedia infoboxes (Suchanek et al., 2007)) or NELL (which extracts facts from any Web text (Carlson et al., 2010; Fader et al., 2011)) are generally static. They are not updated as the Web changes when in reality new facts arise while others cease to be valid or change over time. One approach towards real-time population of KBs is to extract facts between entities from dynamic content of the web such as news, blogs and user comments in social media (Nakashole and Weikum, 2012). This paper proposes a *shift* of focus from doing KB updates by extracting facts in text to doing them by identifying state changes brought about by addition or deletion of state-changing verbs in text.

The benefit of such shift is multi-fold. (1) Even when relation between entities is not stated explicitly in text, detecting state change that happens to an entity in text can be used to infer and update the corresponding relational fact in KB and to temporally scope the fact in KB (Wijaya et al., 2014). (2) Learning state changes brought about by verbs can pave ways to learning the pre- and post-conditions of state-changing verbs: the entry condition (in terms of KB facts) that must be true of the verb's entities for a state-changing event expressed by the verb to take place, and the exit condition (in terms of KB facts) that will be true of the entities after the event occurs. Such pre- and post-conditions can be useful for learning event sequences such as scripts (Schank and Abelson, 2013), that can be modeled as a collection of verbs chained together by the pre- and post-condition overlap of their shared entities, for inferring how the effect of one event can be cascaded down to other entities via the pre- and post-condition of the shared entities, or for inferring unknown states of entities from the verbs they participate in.

In this paper we propose to learn state changes brought about by verbs from the Wikipedia edit

histories of entities. Our assumption is that when a state-changing event happens to an entity e.g., marriage, the Wikipedia infobox that contains the underlying KB facts of the entity is updated e.g., by the addition of a new SPOUSE value. At the same time, texts containing verbs expressing the event e.g., *wed* may be added or deleted from the entity's Wiki page. Wikipedia edits of verbs and infobox over many entities that undergo a similar event can act as weakly supervised data for learning verbs and infobox changes that relate to the event. However, Wikipedia infobox edits are notoriously *noisy*: there is no guarantee that only the infobox slots that relate to the particular event will be updated. For example, when an event such as death happens to an entity, infobox slots regarding the entity's birth e.g., *birthdate*, *birthplace*, may also be updated. To alleviate the effect of such noise, we also leverage constraints between infobox slots e.g., that *deathdate* is mutually exclusive with *birthdate* or that *birthdate* is related to *birthplace*, to effectively learn infobox changes that relate to a particular event-expressing verb. From our experiments, we observe that when verbs expressing a state changing event are being added/deleted to an entity's Wikipedia text, we can update the infobox facts about the entity effectively, with an 89% precision and 74% recall.

2 Method

We construct a dataset from Wikipedia edit histories of person entities whose facts change between the year 2007 and 2012. We consider entities to have facts changed in this period whenever at least one of their facts in Timely YAGO KB (Wang et al., 2010) has begin/end time in this period. Using this method, we obtain Wikipedia URLs of 20,324 entities and crawl their Wikipedia revision histories, obtaining any revisions their Wikipedia pages have between the year 2007 and 2012. Each document in our data set is the *difference* between any two revisions to an entity's Wikipedia page that are separated by at least a single day worth of revisions. For example, a Wikipedia entity "Ralph McInerny" has his page revised on the days of 20 November 2012, 26 and 29 December 2012 consecutively. We find the difference between the first revision on 20 November 2012 and the last revision on 29 December 2012 (since a page can be revised multiple times in a day). This difference, a HTML page obtained by "compare selected revisions" functionality in Wikipedia, is a document in our dataset. By crawling Wikipedia revision histories, we obtain 288,184 documents this way[1].

3 Experiments

4 Related Works

Related works here

5 Conclusion

Acknowledgments

We thank members of the NELL team at CMU for their helpful comments. This research was supported by DARPA under contract number FA8750-13-2-0005.

References

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Ndapandula Nakashole and Gerhard Weikum. 2012. Real-time population of knowledge bases: opportunities and challenges. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 41–45. Association for Computational Linguistics.

Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. 2010. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 697–700. ACM.

[1]We make our Wikipedia edit histories dataset available here: http://www.cs.cmu.edu/~dwijaya/wiki-edits-dataset.zip

Derry Wijaya, Ndapa Nakashole, and Tom Mitchell. 2014. Ctps: Contextual temporal profiles for time scoping facts via entity state change detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.