"A Spouse Begins with a Deletion of *engage* and Ends with an Addition of *divorce*": Mapping Text and Infobox Changes in Wikipedia to Learn Verbs and State Changes for Knowledge Base Updates

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 verbs on its arguments (i.e., entities). State changes are viewed as updates of KB facts pertaining to the entities. We propose to learn state changes brought about by verbs using Wikipedia revision histories. When a state-changing event happens to an entity, the Wikipedia infobox that contains facts of the entity may be updated. At the same time, text that contain verbs that express the event may also be added to or deleted from the entity's Wikipedia page. We use Wikipedia revision histories as distantly supervised data to automatically learn verbs and state changes. We also leverage constraints such as mutual exclusion among infobox slots and simultaneously updated slots to effectively map text to infobox changes. We observe in our experiments that when state-changing verbs are being added to or deleted from 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. One approach towards real-time population of KBs is to extract facts from dynamic content of the web such as news (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 verbs in text.

The benefit of such shift is multi-fold: (1) Detecting state change to an entity in text can be used to infer and update the entity's fact and its temporal scope 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 for an event expressed by the verb to take place, and the exit condition (in terms of KB facts) that will be true after the event occurs. Such pre- and post-conditions can be useful for (a) learning event sequences as a collection of verbs chained together by pre- and post-condition of their shared entities, (b) for inferring cascading effect of an event via the pre- and post-condition of shared entities in an event sequence, or (c) 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 using Wikipedia revision histories. Our assumption is that when a state-changing event happens to an entity e.g., a marriage, its Wikipedia infobox: a structured document that contains a set of facts (attribute-value pairs) of the entity is updated e.g., by the addition of a new SPOUSE value. At the same time, texts that contain verbs that express the event e.g., *wed* may be added to the entity's Wikipedia page (see an example in Figure 1). Wikipedia revisions over many entities can act as distantly supervised data for mapping text and infobox changes that re-

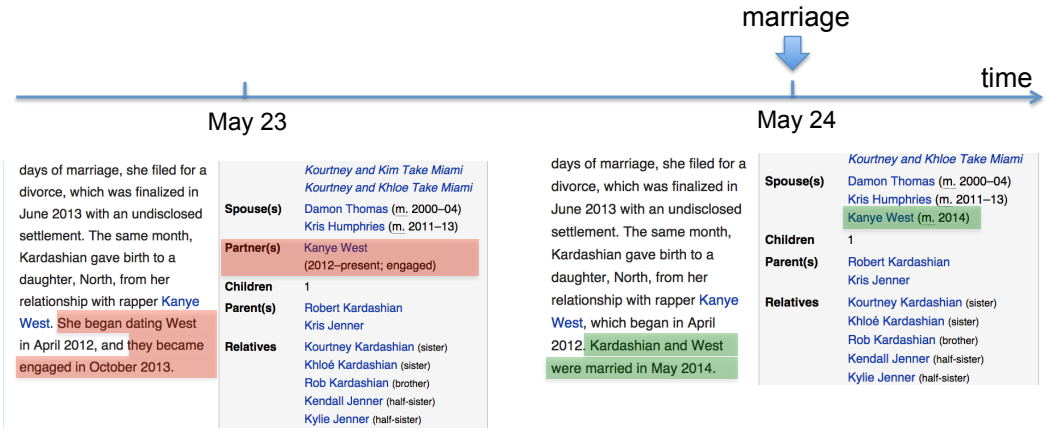

Figure 1: A snapshot of Kim Kardashian's Wikipedia revision history, highlighting text and infobox changes. In red (and green) are the difference between the page on May 23 and May 24, 2014: things that are deleted from (and resp. added to) the page.

late to events. However, Wikipedia revisions are *noisy*: there is no guarantee that only the infobox slots related to a particular event will be updated. For example, when an event such as death happens, slots regarding birth e.g., *birthdate*, *birthplace*, may also be updated. To alleviate the effect of such, we leverage constraints between slots e.g., that *deathdate* is mutually exclusive with *birthdate* or that *birthdate* is simultaneously updated with *birthplace*, to effectively learn infobox changes that relate to a particular event-expressing verb.

Our contribution is the construction and use of an interesting, distantly labeled, dataset from Wikipedia revisions for learning about verb and state changes[1], and the learned resource of verbs effective for identifying state changes.

2 Method

We construct a dataset from Wikipedia revision histories of person entities whose facts change between the year 2007 and 2012. We consider entities to have changed facts whenever at least one of their facts in YAGO KB (Wang et al., 2010) has begin/end time in this period. We obtain Wikipedia URLs of these entities from YAGO and crawl their Wikipedia revision histories, obtaining revisions their 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" page was consecutively revised on the days of 20 November 2012, 26 and 29 December 2012. We find the difference between the first revision to his page on 20 November 2012 and the last revision to his page on 29 December 2012 (since a page can be revised multiple times in a day). This difference[1], a HTML page obtained by "compare selected revisions" functionality in Wikipedia, is a document in our dataset. Using this method, we obtain 288,184 documents in our dataset from revision histories of 16,909 Wikipedia entities[2].

From this dataset, we obtain documents that contain infobox changes. We define an infobox attribute of an entity e.g., SPOUSE to *begin* when a new value or a begin time is added to the attribute slot and to *end* when an end time is being added to the slot. Using regular expression to detect whether a new value, a start, or an end time is being added to infobox slots of a document, we automatically label each document with "begin-{attribute_name}" or "end-{attribute_name}". So a document that contains an addition of a new value in the SPOUSE slot will be labeled "begin-spouse", while a document that contains an addition of end time in the SPOUSE slot will be labeled "end-spouse".

[1]We make our dataset available here: http://.../wiki-edits-dataset.zip

[1]http://en.wikipedia.org/w/index.php?title=Ralph_McInerny&type=revision&diff=530257160&oldid=523980632

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

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.

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.