EVENTS, ENTITIES, AND TEXT

Derry Tanti Wijaya (Carnegie Mellon University)
dwijaya@andrew.cmu.edu

When an event happens in the world, it involves a number of entities, some of which are then influenced or changed by the event (e.g., the person being affected by a shooting event). We have been studying how entities in the world change with respect to how they are described in text.

The series of studies that we conducted ask these questions: (1) can we track changes that happen to entities from the way they are described in text over time? (2) Can we find in the context within which a changed entity is mentioned in text, words that *trigger* the changes that happen to the entity? And are these trigger words indicative of the event that causes the changes? (3) Given the trigger words (we focus on verb phrases) and the changes in text that we learn to relate to a particular event, can we predict when the same event happens to other entity from its description in text? And can we automatically update our knowledge about the entity based on the changes brought about by the event?

We will present key findings from our studies with respect to each of the questions in turn. By presenting these series of studies, we are interested to start a discussion on how such verb- and entity-centric study of events in text can enrich the study of event structure in language.

Tracking changes that happen to entities in text

Tracking the frequency of an entity's mentions in a large collection of texts over time can indicate *when* a change is happening to the entity as its frequency changes [1]. Inspired by this earlier study, we seek to learn not only *when* but also *how* the entity changes from the change in the words that surround the entity over time [2]. We identify clusters of topics surrounding the entity over time and indicate that there is a topic change when two consecutive years are of different clusters.

We use K-means and Topics-over-Time clustering [3]; a Latent Dirichlet allocation (LDA)-style topic model that captures how latent structure in the data changes over time. For example, for the entity *Iran*, our clustering identifies a topic change in the year 1979 with the emergence of a new cluster consisting of words such as 'republic' and 'revolution'. Coincidentally, the year 1979 is the year of the revolution event that changed *Iran* from a monarchy to an Islamic republic. Another example, for the entity *Kennedy*, our clustering identifies a topic change in the year 1961 with a cluster containing words such as 'senator' transitioning to a new cluster containing words such as 'president'. Coincidentally, the year 1961 is the year *John F. Kennedy* was elected president. These findings suggest that we can identify *when* and *how* entities change over time in text and that these changes often coincide with events that happen to the entities.

We also observe that changes happen not only to entities in text but also to words that are adjectives or nouns such as *gay*. For the word *gay* our clustering identifies a topic change in the year 1970 with a cluster consisting of words such as 'happy' and 'lively' transitioning to a cluster with words such as 'lesbian', 'liberation', and 'movement'. Coincidentally, The year 1970 is also the year that the homosexuality movements began to pick up pace.

Finding words that trigger entities' change in text

In the following study [4], we seek to discover words that trigger entities' change in text and to find if these words are indicative of the event that causes the change. We start with some example entities; we call them *seed* entities that undergo a similar event at some points in in their lives. We then aggregate the context of these seeds, which means averaging the frequency of words within which these 'seed' entities are mentioned in text: *before*, and *at* the time of the event.

Our hypothesis is that the context (surrounding words) *at* the time of the event contains words that are indicative of the event while the *difference* in contexts *before* and *at* the time of the event reflects changes associated with the event. Our observation confirms this hypothesis. For example, the aggregate context of seed entities for US presidency election contains words such as 'was elected', 'took office' which are indicative of the election event. We also observe that the *difference* in contexts before and after the event reflects the changes associated with the election event with the rise in mentions of 'administration' and 'president' surrounding the entity and the fall in mentions of 'senator', 'governor' and 'candidate' surrounding the entity. These findings suggest that from the frequency words surrounding the entities in text we can identify trigger words that are indicative of the event that causes the entities' changes.

Predicting change in entities

In this last study [5] we come full circle. We use trigger words that indicate a particular event to predict the change in the entities that is brought about by the event.

We focus on trigger words in the form of verb phrases that indicate events and the changes associated with the events. We learn these verbs and the associated changes from Wikipedia edit history. Our motivation is as follows: when an event happens to an entity, the *infobox* on the entity's Wikipedia page usually gets updated. At the same time, the article text may be updated with verbs either being added or deleted to reflect changes made to the infobox.

We use Wikipedia edit history to distantly supervise a Maximum Entropy classifier for automatically learning verbs and the associated changes that they bring. Our classifier learns that verb phrases such as 'die on' or 'pass on' trigger the deaths of their subjects (i.e., death dates are being added to their infoboxes). Verb phrases such as 'marry on' trigger their subjects to 'gain' a spouse (i.e., spousal information is being added to their infoboxes) while verb phrases such as 'divorce in' trigger the subjects to 'lose' a spouse. From the experiments, we observe that when these trigger verbs are being added or deleted from an entity's Wikipedia page, we can predict the entity's infobox changes with 88% precision and 76% recall. These findings suggest that we can automatically discover verb phrases in text that reflect events and that these verb phrases are effective for predicting entities' changes caused by the events.

References

[1] Michel, Jean-Baptiste, et al. "Quantitative analysis of culture using millions of digitized books." science 331.6014 (2011): 176-182.

[2] Wijaya, Derry Tanti, and Reyyan Yeniterzi. "Understanding semantic change of words over centuries." *DETecting and Exploiting Cultural diversiTy on the social web workshop*. ACM, 2011.

[3] Wang, Xuerui, and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

[4] Wijaya, Derry Tanti, Ndapandula Nakashole, and Tom M. Mitchell. "CTPs: Contextual Temporal Profiles for Time Scoping Facts using State Change Detection." *EMNLP*. Doha, Qatar. 2014.

[5] Wijaya, Derry Tanti, Ndapandula Nakashole, and Tom M. Mitchell. ""A Spousal Relation Begins with a Deletion of engage and Ends with an Addition of divorce": *EMNLP*. Lisbon, Portugal. 2015.