Learning State Change Sequences from Temporal Profiles of Entities to Detect Knowledge Base Updates

Abstract

Methods for information extraction (IE) and knowledge base (KB) construction have been widely studied in recent years. However, a largely under-explored case is maintenance of knowledge once it has been acquired. Attributes of entities in the knowledge base change over time. Capturing such state changes has implications for the correctness of the KB. Current IE methods are not capable of detecting such changes. In this paper, we present a method for detecting state changes in attributes values of KB entities. Our method is based on a identifying state change sequences over temporal profiles of similar entities. Our experiments show the potential of a our approach.

1 Introduction

Motivation. Recent progress in automatic knowledge acquisition has resulted in a number of large knowledge bases (KBs) (Bollacker 2008; Carlson 2010; Suchanek 2007) . Such KBs contain many millions of entities, organized in hundreds to hundred thousands of semantic classes, and hundred millions of relational facts between entities. With this progress comes new research problems regarding maintenance of KBs. One such problem is the detection of state changes to entities. When one attribute value for a given entity is no longer true, we need to update the knowledge base. This can occur for example if a person gets divorced, they are no longer the spouse of who they were married to. Furthermore, when someone is fired or resigns from their job, they are no longer employees of their current employer. Currently, most IE methods detect patterns for learning attributes and mix them together with those involving state changes in the attribute. It is not unusual for an IE system to learn that the phrases: "is married to" and "is divorced from" from are both good for indicating the "hasSpouse" attribute. In reality, one of those phrases marks the beginning and the other marks the end of an attribute value for the "hasSpouse" attribute.

Problem Statement. Prevalent approaches to IE extract knowledge from static Web snapshots such as the ClueWeb crawl[1] (Fader 2011; Nakashole 2011), with no mention of how to perform updates. Other approaches periodically extract from a corpus such as Wikipedia, every time re-applying the extractor to all the documents even those that did not change (Suchanek 2007). The NELL system (Carlson 2010) follows a never-ending' extraction model with the extraction process going on 24 hours a day. However NELL's focus is on language learning to self improve on its reading ability over time. In contrast, here we focus on detecting updates to specific attributes of entities. Detecting these changes has unique challenges not seen in IE methods.

1. **Finer grained language understanding :** Learning state changes requires differentiating between language used to indicate different states of a given attribute. This is a much harder task than learning phrases that are related to a given attribute in some *any* way.

2. **Targeted extraction model:** IE methods often extract everything they are able to find in a given corpus. However, to capture state changes, we need to instead employ a targeted extraction model. This model needs to find documents that are promising for state change information for a given entity. A method that identifies the documents of inter-

[1] lemurproject.org/clueweb09.php/

est can leverage news aggregation and micro news platforms such as Twitter.

In this work we focus on the first challenge of finer grained language understanding and leave the second one for future work.

Contribution and Paper Organization. We developed a method for detected state changes. Our approach finds state change sequences across temporal profiles of similar entities. TO DO: describe method fully, followed by overview of the rest of the paper.

References

A. Angel, N. Koudas, N. Sarkas, D. Srivastava: Dense Subgraph Maintenance under Streaming Edge Weight Updates for Real-time Story Identification. In *Proceedings of the VLDB Endowment*, PVLDB 5(10):574–585, 2012.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z.G. Ives: DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, pages 722–735, Busan, Korea, 2007.

M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni: Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2670–2676, Hyderabad, India, 2007.

K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor: Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages, 1247-1250, Vancouver, BC, Canada, 2008.

A. Carlson, J. Betteridge, R.C. Wang, E.R. Hruschka, T.M. Mitchell: Coupled Semi-supervised Learning for Information Extraction. In *Proceedings of the Third International Conference on Web Search and Web Data Mining (WSDM)*, pages 101–110, New York, NY, USA, 2010.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr., T. M. Mitchell: Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI)* 2010.

L. Del Corro, R. Gemulla: ClausIE: clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 355-366. 2013.

A. Das Sarma, A. Jain, C. Yu: Dynamic Relationship and Event Discovery. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining (WSDM)*, pages 207–216, Hong Kong, China, 2011.

A. Fader, S. Soderland, O. Etzioni: Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1535–1545, Edinburgh, UK, 2011.

C. Havasi, R. Speer, J. Alonso. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2007.

J. Hoffart, F. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, G. Weikum: YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 229–232, Hyderabad, India. 2011.

Karin Kipper, Anna Korhonen, Neville Ryant, Martha Palmer, A Large-scale Classification of English Verbs, Language Resources and Evaluation Journal, 42(1): 21-40, 2008, data available at `http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html`

N. Nakashole, M. Theobald, G. Weikum: Scalable Knowledge Harvesting with High Precision and High Recall. In *Proceedings of the 4th International Conference on Web Search and Web Data Mining (WSDM)*, pages 227–326, Hong Kong, China, 2011.

N. Nakashole, T. Tylenda, G. Weikum Fine-grained Semantic Typing of Emerging Entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1488-1497, 2013.

N. Nakashole, G. Weikum, F. Suchanek: PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1135 - 1145, Jeju, South Korea, 2012.

Feng Niu, Ce Zhang, Christopher Re, Jude W. Shavlik: DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. In the VLDS Workshop, pages 25-28, 2012.

A. Ritter, Mausam, O. Etzioni, S. Clark: Open Domain Event Extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1104-1112, Beijing, China, 2012.

D. Shahaf, Carlos Guestrin: Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 623-632, 2010.

T. Sakaki, M. Okazaki, Y. Matsuo: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 851-860, Raleigh, North Carolina, USA, 2010.

F. M. Suchanek, G. Kasneci, G. Weikum: Yago: a Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW)* pages, 697-706, Banff, Alberta, Canada, 2007.

F. M. Suchanek, M. Sozio, G. Weikum: SOFIE: A Self-organizing Framework for Information Extraction. In*Proceedings of the 18th International Conference on World Wide Web (WWW)*, pages 631–640, Madrid, Spain, 2009.

W. Wu, H. Li, H. Wang, K. Zhu: Probase: A Probabilistic Taxonomy for Text Understanding. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 481–492, Scottsdale, AZ, USA, 2012.

Word Wide Web Consortium (W3C): RDF Primer. `http://www.w3.org/TR/rdf-primer/`, Accessed 2013.