4K1-OS-2-5

Webからの関係抽出とそれを利用した関係検索

Danushka Bollegala[*1]

[*1]The University of Tokyo

In this talk, I will introduce the problem of extracting semantic relations between entities from the Web. First, I will introduce the prior research work in this area and then explain potential applications of relation extraction. The talk is intended as an overview in Web-based semantic relation extraction. This supplementary paper contains related references that will be useful for the interested participants to obtain further technical details.

1. Semantic Relations between Entities

Entities and relations are omnipresent in Web data. Web texts such as online newspaper articles, blog posts, encyclopedic resources such as Wikipedia[*1], online social networks such as Facebook[*2] contain numerous types of named entities such as people, organizations, locations, and products. Entities are often closely connected to other entities in the Web via numerous semantic relations. For example, one company might acquire another company or one person might get married to another. Organizing the information in the Web as a network of inter-connected entities via semantic relations is useful for both visualization purposes as well as efficient information retrieval. As a concrete example, the online social search system SPYSEE[*3] organizes people based on typical semantic relations that exist between people.

Correct identification of semantic relations between entities is an important first step in numerous tasks. For example, in Web Information Retrieval [18], if a user issues a query regarding a particular person, then we can return results not only about that person but also about his or her place of work, colleagues etc. to improve the search experience. However, extracting semantic relations from web texts is a challenging task due to several reasons. First, entity resolution is a difficult problem in the Web. Multiple entities such as people might have the same name (i.e. namesake disambiguation problem) [3] as well as a single entity might be referred to by multiple names (i.e. name alias detection problem) [6]. For example, for the name *Jim Clarke*, we find numerous results on the Web for both the late F1 racing champion as well as for the Netscape founder. On the other hand, *Will Smith* is popularly known as the *fresh prince* in Web contexts. If we cannot correctly resolve the entities, then it becomes impossible to correctly extract relations among those entities. Second, there can be multiple semantic relations between two given entities [4]. For example, ostrich *is the largest* bird as well as *a flightless* bird. Third, a semantic relation can be expressed in numerous ways – both **X** *acquired* **Y** as well as **X** *purchased* **Y** indicate an ACQUISITION relation between **X** and **Y**. To identify relations, we must correctly map the different paraphrases to each relation. The term *relation type* is often used to refer to a particular semantic relation (e.g. ACQUISITION) whereas, the term *relation instance* refers to a particular instance of a relation type (e.g. (Google, YouTube) is an instance of ACQUSITION).

2. Problem Settings and Approaches

Relation extraction can be broadly classified into two categories: sentence-level and corpus-level. In sentence-level relation extraction [7, 20, 19], we are required to determine whether a particular semantic relation exists between two given entities. Therefore, it can be seen as a binary classification problem in which we are required to classify a given sentence as positive only if the relation R holds between the two entities e_i and e_j in a sentence s_k. In domains where the set of relation types that we are interested in extracting is pre-specified, we can use labeled training data to train a binary classifier for each relation type or alternatively train a single multi-class classifier for all relation types. However, in the case of a single multi-class relational classifier, we must first filter-out cases where there is no relation between the two entities.

On the other hand, in the corpus-level relation extraction setting, we are given a text corpus and are required to extract *all* relation instances that exist between all pairs of entities. This setting is also known as Open Information Extraction (Open IE) [21, 2, 17, 1, 9, 8, 14]. In this setting, the set of relation types is not given in advance. Open IE can be seen as an unsupervised semantic relation extraction task that closely models the relation extraction scenario encountered on the Web [5]. However, this is a challenging setting and some form of a supervision is often required to improve the accuracy of the relation extraction. One popular method of supervision is to provide a small number of seed instances of the relation type that we are interested in extracting. In addition, we can also provide a few extraction patterns for the relation type. Both seed instances as well as seed patterns are then used in a bootstrapping algorithm to extract more relation instances for the relation type [15]. This approach is often referred to as *semi-supervision*, *weak-supervision* or *distant supervision* in the literature. However, it has been shown that after few

連絡先: 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
danushka@iba.t.u-tokyo.ac.jp

[*1] http://www.wikipedia.org/

[*2] http://www.facebook.com/

[*3] http://spysee.jp/

iterations of the bootstrapping process, there is a risk of extracting relation instances that are not related to the seeds. This phenomenon is called the *semantic drift*, and control measures must be taken to avoid it in practice. Moreover, the bootstrapping process is shown to be sensitive to the initial seeds [13].

3. Applications

Relation extraction has been successfully used in numerous tasks such as information retrieval, paraphrase extraction, and social network mining. Next, I will describe several use cases of relation extraction.

Latent Relational Search[*4] is a novel search paradigm that focuses on the implicit semantic relations between two entity pairs [11]. Given three entities A, B, and C, in latent relational search we retrieve entities D for which the semantic relation between A and B is similar to that between C and D. For example, given the relational search query (*U.S.*, *Lady Gaga*), (*Japan*, ?), we would like to retrieve entities that have similar semantic relations to Japan as the semantic relations between U.S. and Lady Gaga. Latent relational search is useful when we cannot explicitly state the relation that we would like to search. In the previous example, there can be various semantic relations between Lady Gaga and U.S. such as Lady Gaga being a U.S. singer, an fashion idol, and a philanthropist. As an interesting extension of the latent relational search paradigm, Duc et al. [10] proposed cross-lingual latent relational search. In this setting, the two entities A and B in a relational search query is given in one (source) language, and the third entity C is given in a different (target) language. For example, a cross-lingual latent relational search engine can be used to answer the query (*U.S.*, *Barak Obama*), (日本,?). Here, the cross-lingual latent relational search engine must return entity names in Japanese that has similar relations with 日本 (Japan) as to Barak Obama with U.S.

Relation extraction is useful for visualization and navigation in a large corpus. We can perform vertical searches along one or more semantic relations. Entity Cube[*5] by Microsoft Research is an interesting tool that enables a user to visualize and search information on entities using the semantic relations among those entities. Considering the vast amounts of textual data available electronically, such visualization and navigation techniques will be important in the near future.

Social networks such as Facebook has gained tremendous popularity over the last few years. When a new user joins an online social networking system, initially however that user will have fewer number of connections (friends). With insufficient link information it is difficult to accurately recommend friends to new users. This scenario is often referred to as the cold start problem. Relation extraction can be used to find people related to a particular person such as friends, colleagues, co-authors, etc. from Web texts for recommendation purposes.

4. Future Research Directions

Despite the numerous potential applications of relation extraction and the research work conducted in the natural language processing field, there are many unsolved issues in relation extraction. Two important research directions are discussed below.

Much research in relation extraction focus on binomial relations that involve two entities. Although most semantic relations can be considered as predicates with two arguments, there are semantic relations that involve more than two entities. For example, in Twitter[*6], user A might *favourite* a *tweet* T posted by another user B. Therefore, the relation FAVOURITES involves all three entities A, B, and T. Although it is possible in principle to decompose a multinomial relation into a set of binomial relations, in doing so we loose the important constraint that all arguments originally satisfied the same predicate. Methods that can represent and extract variable size multinomial relations without breaking them into binomial relations must be studied in the future.

Similar to entities, relations can also be ambiguous. Ambiguity in semantic relations can appear at multiple complexity levels. First, there is lexical ambiguity in which single lexical pattern can subsume multiple semantic relations. This type of ambiguity is closely related to the Word Sense Disambiguation (WSD) problem. Second, given a set of n entities we must test a maximum of $n(n-1)/2$ number of pairs (for binomial relations) for the existence of a semantic relation. Trying all pairs in large datasets such as the Web is infeasible and some sort of locality constraints must be imposed. In the extreme, we can limit relation extraction to entities in the same sentence. These constrains can be relaxed if we consider co-reference chains in texts [12]. This referential ambiguity in semantic relations must be carefully studied in future research to both improve the coverage of relation extraction as well as to improve its scalability.

5. Semantic Relations and AI

Building an AI system requires two fundamental tasks: *knowledge representation* and *inference*. Both those tasks can be seen as relation extraction tasks. Classical knowledge representation methods such as first order logic represents a knowledge base as a set of deterministic rules which consists of predicates and arguments. Arguments can be seen as entities, and predicates corresponds to semantic relations. Therefore, relation extraction from raw text can be considered as the task of extracting the knowledge base required by an AI system. Next, the inference process can also be seen as a one of relation detection between existing knowledge and newly encountered fact that must be verified. If a series of relations can be found that connects one or more relational predicates in the existing knowledge base to the newly encountered fact, then we can say that we can *infer* the newly encountered fact from our existing knowledge base. By replacing deterministic rules with prob-

[*4] http://www.milresh.com/

[*5] http://entitycube.research.microsoft.com/

[*6] https://twitter.com/

abilistic ones and by conducting probabilistic inference, it is possible to model the real-world information more accurately [16].

参考文献

[1] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI'07*, pages 2670–2676, 2007.

[2] M. Banko and O. Etzioni. The tradeoffs between traditional and open relation extraction. In *ACL'08*, pages 28–36, 2008.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka. Extracting key phrases to disambiguate personal name queries in web search. In *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval*, pages 17–24, 2006.

[4] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Measuring the similarity between implicit semantic relations from the web. In *WWW 2009*, pages 651 – 660, 2009.

[5] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *WWW 2010*, pages 151 – 160, 2010.

[6] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Automatic discovery of personal name aliases from the web. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 23(7):977 – 990, July 2011.

[7] R. Bunescu and R. Mooney. Subsequence kernels for relation extraction. In *NIPS'06*, pages 171–178, 2006.

[8] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *IJCAI'05*, pages 659–664, 2005.

[9] D. Davidov, A. Rapport, and M. Koppel. Fully unsupervised discovery of concept-specific relationships by web mining. In *Proc. of ACL'07*, pages 232–239, 2007.

[10] Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka. Cross-language latent relational search: Mapping knowledge across languages. In *AAAI'11*, pages 1237 – 1242, 2011.

[11] Nugyen Duc, Danushka Bollegala, and Mitsuru Ishizuka. Using relational similarity between word pairs for latent relational search on the web. In *Proc. of Int'l Conf. on Web Intelligence (WI'10)*, 2010.

[12] Ryan Gabbard, Marjorie Freedman, and Ralph Weischedel. Coreference for learning to extract relations: Yes, virginia, coreference matters. In *ACL'11*, pages 288 – 293, 2011.

[13] Zornista Kozareva and Eduard Hovy. Not all seeds are equal: Measuring the quality of text mining seeds. In *NAACL-HLT 2010*, 2010.

[14] D. Lin and P. Pantel. Dirt: Discovery of inference rules from text. In *Proc. of ACM SIGKDD'01*, pages 323–328, 2001.

[15] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *ACL'06*, pages 113 – 120, 2006.

[16] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62:107 – 136, 2006.

[17] B. Rosenfeld and R. Feldman. Clustering for unsupervised relation identification. In *In proc. of CIKM'07*, pages 411–418, 2007.

[18] G. Salton and C. Buckley. *Introduction to Modern Information Retreival*. McGraw-Hill Book Company, 1983.

[19] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *JMLR*, 3:1083–1106, 2003.

[20] G. Zhou, M. Zhang, D. H. Ji, and Q. Zhu. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *EMNLP-CoNLL*, pages 728 – 736, 2005.

[21] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji R. Wen. Statsnowball: a statistical approach to extracting entity relationships. In *WWW'09*, pages 101–110. ACM, 2009.