TLAB at the NTCIR-13 AKG Task

Md Mostafizur Rahman
Sokendai(Graduate University for Advanced Studies)
National Institute of Informatics
Tokyo, Japan
rahman@nii.ac.jp

Atsuhiro Takasu
National Institute of Informatics
Tokyo, Japan
takasu@nii.ac.jp

ABSTRACT

In recent years, popular search engines are utilizing the power of Knowledge Graph(KG) to provide specific answers to queries and questions in a direct way. It is expected that search engine result pages (SERPs) will provide facts about the quires satisfying semantic meaning, which encouraging researchers to constructing more powerful Knowledge Graph. One of the major challenges is disambiguating and recognizing entities and their actions stored in KG in a context. To achieve and advance the technologies related to actionable knowledge graph presentation, Action Mining (AM) is an essential step and relatively new research direction to nurture research on generating such KG that is optimized for facilitating entity's actions e.g. for entity "Donald J. Trump" most potential actions could be "won the US Presidential Election" or "targeting US journalists". This paper presents the Action Mining (AM) task organized by NTCIR-13. We employ a probabilistic model to address the AM problem.

Team Name

TLAB

Subtasks

Action Mining (AM)

Keywords

Entity, Actions, Action Mining

1. INTRODUCTION

Data is generating from everywhere around us all the times. Smart phones, sensors and social networking sites produce tons of data everyday. In recent years, large amount of data is being available on web so handling abundance of information and extract facts can be considered as major challenges of search engines. Most of the people consider search engines as an expert of all domain. As the expectation is heading towards peak, effective use of KG in SERPs becomes essential. To achieve such goals, it is mandatory to design more structured and sophisticated knowledge graph generation technique.

Large Knowledge Bases (KB) e.g. DBpedia, YAGO, DeepDive are incorporating huge number of entities and attributes to keep pace with the high information generation in web, smart systems and social life. Evolving with new facts, at the same time organizing and maintaining existing knowledge is more critical.

The purpose of AKG defined by NTCIR-13[1] [3] is: select and rank attributes of entities in KGs that can best support "actionable" search intents. To prepare a KG for supporting actionable search two major steps are: (1) Recommend the actions relevant to queries, (2) Ranking the attributes of query entity based on action and graph generation. This study presents, action mining technique to foster the actionable knowledge graph generation. In short, main goal is to find the top actions relevant to query entity and entity type and embed the entities with their related counterparts.

To the best of our knowledge, there is no existing work on the entity oriented action mining for actionable knowledge graph generation. Related entity mining and semantic role labelling (SRL) can be considered as the similar topics to the problem presented in this paper. In related entity recommendation, goal is to retrieve top related entities given a keyword query. Web search engines more often use their own data gathered from users as well as user click logs and sessions to recommend related entities [4, 12, 1]. In this study, we employ mostly publicly available data to generate a list of top actions relevant to query. In SRL, predicates or verbs use for detection of the semantic arguments and identify the role of entity [5]. In AM problem based on entity we recommend the top related action that match with query entity and entity type. We propose a simple but statistically sound probabilistic model and discuss the parameter estimation of the model.

2. PROBLEM STATEMENT

In this section, apart from the NTCIR description, we formally define the action mining problem with some examples and briefly describe our approache to address the AM problem.

2.1 Problem Definition

In entity oriented action mining problem, the goal is to find the potential actions and return *top-k* potential actions relevant to query, where query is a set of entity instance and entity type [3].

According to English grammatical rule, subject, object and verb form a sentence. The subject and the verb are the minimum requirements for constructing a basic English sentence. Verb plays the key role to give semantic meaning to a sentence. As we know that an auxiliary verb[2] is used in forming tense and a linking verb[3] joins the the subject with

[1] http://research.nii.ac.jp/ntcir/
[2] https://en.wikipedia.org/wiki/Auxiliary_verb
[3] https://en.wikipedia.org/wiki/Linking_verb

Table 1: Input and Output of Action MiningTask.

Entity Instance	Entity type	Action Verb	Object
Madrid	Place	represent	European arts and culture
		visit	on holiday
		recommend	some good restaurants
Google	Organization	make	huge revenue
		buy	DeepMind Lab
		prepare	for a phone interview with google

complement. Unlike auxiliary or linking verb, an action verb expresses physical, objective, mental or psychological action of a subject or entity in a sentence. For example, "Mozart could play the piano blindfolded", here "play" is an action verb representing a physical action of entity "Mozart" and "could" is used as an auxiliary verb to form the tense correctly. So, an action consists of action verb and associated object (if any). To get a clear view of the AM problem, please refer to Table 2. Here, we define action and entity oriented action mining problem formally.

DEFINITION 1. ***(Action)***

An action a, comprises with action verb v_a and associated object o_a (if any).

DEFINITION 2. ***(Action Mining)***

Given text data sources, the input is a query q, which is a set of entity instance e_q and entity type t_q, and the output is a list of top-k actions relevant to query q.

2.2 Our Approach

We propose a Probabilistic Model for Action Mining (PMAM). In our model we decompose the AM problem into several distributions which reflect heterogeneous relationship between query and relevant actions including popularity, entity-action relatedness and entity type-action relatedness. The goal is estimating $P(a|q)$ or $P(a|e_q, t_q)$ of each action a, given query entity instance e_q and entity type t_q. We describe the components of $P(a|e_q, t_q)$ in Section 4.

This paper discusses action mining problem which is quite new topic. The idea of action mining has some similarity with entity recommendations. Sundog [4] and Spark [2] are proposed by Yahoo! for related entity recommendations in web search, exploit supervised learning techniques. On the other hand, Microsoft proposed Three-way Entity Model (TEM) [1] that provides personalized recommendation of related entities, which is basically a probabilistic approach. Here, we choose the probabilistic approach.

3. DATA EXTRACTION

In this section, we describe data cleaning, selection and extraction process of our system. Raw data can be very noisy, inconsistent and incomplete. For each group of data, we clean the data, prune the irrelevant data by defining data pruning rules and eventually extract the data and create a database to support our queries.

We remove all kinds of tags and other unnecessary symbols/characters. For data selection and pruning we define two rules:

Rule 1: *Extracted sentences must contain an entity, e and at least one action verb, v_a as defined in problem definition section.*

This rule ensures that extracted sentences only contain actions and other data is simply discarded.

Rule 2: *To handle duplication of same action, lemmatize the action verbs of extracted sentences.*

Lemmatization of a verb means change the verb in base form [7]. Same action can appear in a different form. Please notice each sentence given below:

1. Donald J. Trump wins the election.
2. Donald J. Trump won the eleetion.
3. Donald J. Trump is wining election.

In the above sentences, the action verb "win" appears in several inflected forms, but all of them represent the same action of entity "Donald J. Trump". The purpose of this rule to handle such duplication and consider only base form (win) of any action verb. For these sentences our method will consider only one action of entity "Donald J. Trump": <"win", "election">.

4. PROBABILISTIC MODEL FOR ACTION MINING (PMAM)

In this part, we formalize the problem in a probabilistic model. The goal is estimating $P(a|q)$, for each query q and return top-k actions while maximizing the probability of $P(a|q)$. Based on the Bayes' theorem, the probability of $P(a|q)$ is derived as follows:

$$\begin{aligned} P(a|q) &= \frac{P(a,q)}{P(q)} \\ &= \frac{P(a,e_q,t_q)}{P(q)} \\ &= \frac{P(a)P(e_q|a)P(t_q|a)}{P(q)} \end{aligned} \tag{1}$$

Here, the denominator can be ignored as it has no influence on action ranking. So, we can rewrite it as below:

$$P(a|q) \propto P(a)P(e_q|a)P(t_q|a) \tag{2}$$

The estimation of the components of $P(a|e_q, t_q)$ are derived in the following subsection separately and later joined the full model.

4.1 Popularity Model, $P(a)$

In this model we simply count the frequency of actions to analyze and understand the common distribution of the actions on the data sources.

$$P(a) = \frac{f(a)}{\sum_{a_i} f(a_i)} \tag{3}$$

$f(a)$ is the frequency of action a.

4.2 Entity Relatedness, $P(e_q|a)$

Entity relatedness model investigates the relatedness of entity instance and associated actions. For the co-occurrence of the action and the entity , point-wise mutual information (PMI) is employed.

$$P(e_q|a) = \frac{PMI(e_q, a)}{\sum_{e_q^{(i)} \in \mathcal{E}} PMI(e_q^{(i)}, a)} \tag{4}$$

where $\mathcal{E}$ is set of all entity instances and we calculate PMI as below:

$$PMI(e_q, a) = \log \frac{P(e_q, a)}{P(e_q)P(a)} \tag{5}$$

where $P(e_q) = \frac{f(e_q)}{N}$, $P(a) = \frac{f(a)}{N}$, $P(e_q, a) = \frac{f(e_q, a)}{N}$, $f(e_q)$ is the frequency of sentences contain entity instance e_q, $f(a)$ is the frequency of sentences contain action a, $f(e_q, a)$ is the frequency of sentences where entity e_q co-occur with action a and N is total number of sentences in corpus.

4.3 Type Relatedness, $P(t_q|a)$

Here, we investigate how often action, a appears in sentences depend on entity type, t_q

$$P(t_q|a) = \frac{f(t_q, a)}{\sum_{t_q^{(i)} \in \mathcal{T}} f(t_q^{(i)}, a)} \tag{6}$$

where $f(t_q, a)$ is the frequency of sentences in which a and t_q co-occur and $\mathcal{T}$ is set of all entity types.

5. EVALUATION

In this section, we discuss the data sources and results evaluated by NTCIR-13.

5.1 Data Sources

In our experiment, we employ multiple reliable data sources. To deal with various type of queries, we need to use various type of data e.g., news data, wikidumps, product review, etc.

5.1.1 *Reuters Corpus.*

Reuters Ltd offers a large collections of news articles for research purpose. This paper uses Reuters Corpus Volume I (RCV1) [8], which contains English news stories from 1996-08-20 to 1997-08-19. RCV1 is an archive of over 800,000 manually categorized newswire stories and the corpus size is about 2.6 GB.

5.1.2 *Wikipedia.*

Wikipedia[4] is very well organized and rich data source. Wikipedia resources are openly available and Wikipedia provides text of all pages, page links, media meta data etc. We collect randomly chosen English article pages from wiki dumps[5].

5.1.3 *Leipzig Corpora*

Leipzig Corpora contain open source data collected from newspapers and web resources [6]. It offers data in more than 200 languages. We gather the news data (6M sentences) in 2012, 2013 and 2015 from these well known corpora.

5.1.4 *Trip Advisor and Amazon User Reviews.*

As there might be some queries regarding place and product entities, so we collect chunks of Trip Advisors and Amazon reviews data provided by [11]. From this dataset we only consider the users' reviews.

5.1.5 *Medline Journals.*

Query may contain medical entities. We observed that Wikipedia and news data contain very few information about medical entities e.g. "Allergy" , "Spinal muscular atrophy". So, we explore the Medline dataset to get rich information about medical entities. Medline dataset is a well-known bibliographic database of life sciences and biomedical information. We are only interested in Medline journals.

5.1.6 *Movie Reviews.*

Beside news data and Wikipedia pages, movie reviews are good sources to collect actions regarding "Movie" entities. From this point of view, we gather movie reviews from Large Movie Review Dataset [9] and Movie Review Data [10].

5.2 Evalution of Submissions by NTCIR

NTCIR-13 set up action mining task and participants had been asked to submit relevant actions based on their given query (query consists of entity instance and entity type) set. They collected the actions for their queries from the participants in their designed task. They evaluated the submitted data by crowdsourcing and finally prepared ground truth dataset. For each pair of action and query from the pooled data, the annotators had to choose among the following options:

- L3: Some people, organizations or other subjects definitely have taken or will take this action for the entity.
- L2: This action has been or will be definitely taken by the entity.
- L1: This action can be relevant for the entity.
- L0: There is no relevance of the action to the entity.

5.3 Results

NTCIR employed Normalized Discounted Cumulative (nDCG) and Expected reciprocal rank (nERR)as performance metric with cut-off level k. nDCG is calculated as below:

$$nDCG@k = \frac{DCG@k}{lDCG@k} \tag{7}$$

where

$$DCG@k = \sum_{i=0}^{k} \frac{rel_i}{log_2(i+1)} \tag{8}$$

and $lDCG@k$ is the maximum attainable DCG, rel_i is the graded relevance assigned to the result at position i. We have conducted experiments with varying k and evaluated the performance of the models.

Table 2 shows the results of the 1st and the 2nd assessment (Verb Only and Verb+Modifier) of AM subtask based on our submission, evaluated by the task organizers. In section 5.2, we described the evaluaton process conducted by the organizers.

[4]https://www.wikipedia.org/
[5]https://dumps.wikimedia.org/

Table 2: Results of AM task using verb only and verb+modifier for evaluation

	nDCG@10	nDCG@20	nERR@10	nERR@20
Verb Only	0.6424	0.7549	0.6831	0.6854
Verb+Modifier	0.3577	0.4325	0.3272	0.3358

6. CONCLUSIONS

In this paper, we presented action mining task organized by NTCIR-13 which is a very interesting and challenging as well from the Information Retrieval (IR) perspective. We employed Probabilistic Model for Actions Mining (PMAM). Our model do not rely on specific data sources and can handle heterogeneous data.

7. REFERENCES

[1] B. Bi, H. Ma, B.-J. P. Hsu, W. Chu, K. Wang, and J. Cho. Learning to recommend related entities to search users. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 139–148. ACM, 2015.

[2] R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *International Semantic Web Conference*, pages 33–48. Springer, 2013.

[3] R. Blanco, H. Joho, A. Jatowt, H. Yu, and S. Yamamoto. Overview of ntcir-13 actionable knowledge graph (akg) task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[4] L. Fischer, R. Blanco, P. Mika, and A. Bernstein. Timely semantics: a study of a stream-based ranking system for entity relationships. In *International Semantic Web Conference*, pages 429–445. Springer, 2015.

[5] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[6] D. Goldhahn, T. Eckart, and U. Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, pages 759–765, 2012.

[7] M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[8] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[10] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

[11] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM, 2011.

[12] X. Yu, H. Ma, B.-J. P. Hsu, and J. Han. On building entity recommender systems using user click log and freebase knowledge. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 263–272. ACM, 2014.