KIEM: A Knowledge Graph based Method to Identify Entity Morphs

Longtao Huang
Institute of Information Engineering
Chinese Academy of Sciences
China
huanglongtao@iie.ac.cn

Lin Zhao
Institute of Information Engineering
Chinese Academy of Sciences
China
zhaolin@iie.ac.cn

Shangwen Lv
Institute of Information Engineering
Chinese Academy of Sciences
China
lvshangwen@iie.ac.cn

Fangzhou Lu
Institute of Information Engineering
Chinese Academy of Sciences
China
lufangzhou@iie.ac.cn

Yue Zhai
Beijing Information Science and
Technology University
China
yvettezhai@mail.bistu.edu.cn

Songlin Hu
Institute of Information Engineering
Chinese Academy of Sciences
China
husonglin@iie.ac.cn

ABSTRACT

An entity on the web can be referred by numerous morphs that are always ambiguous, implicit and informal, which makes it challenging to accurately identify all the morphs corresponding to a specific entity. In this paper, we introduce a novel method based on knowledge graph, which takes advantage of both knowledge reasoning and statistic learning. First, we present a model to build a knowledge graph for the given entity. The knowledge graph integrates the fragmented knowledge on how humans create morphs. Then, the candidate morphs are generated based on the rules summarized from the knowledge graph. At last, we use a classification method to filter the useless candidates and identify the target morphs. The experiments conducted on real world dataset demonstrate efficiency of our proposed method in terms of precision and recall.

KEYWORDS

Entity Morphs; Knowledge Graph; Web Mining; Language Understanding

1 INTRODUCTION

Motivation: Information morph is a phenomenon that Internet users create some alias to hide the original objects for various reasons [1]. For example, some outlaws might use morphs to spread malicious information about sensitive people and events in order to evade censorship [2]. Besides, morphs are widely used in online social networks to express users' irony/positive/negative sentiment and make their posts more vivid. Here is an example morph in Chinese social media. “奥胖 (O Fat)” refers to the famous NBA star “*Shaquille O´Neal*” due to his big size and heavy weight. Note that this paper mainly talks about the person entities since most morphs on the web are for persons.

However, the popularity of entity morphs makes it difficult to retrieve all the information about a specific person if we only use his real name. Identifying entity morphs can not only help to find more related information for a person retrieval task, but is also critical for language understanding tasks such as entity linking and event argument extraction [3]. Thus, this paper aims at identifying entity morphs on the web with both high precision and recall.

Problem Statement: This paper addresses the problem of identifying entity morphs corresponding to a given person. The raw input for this task is a set of real names which refer to specific people and all the available information on the web, such as news articles, Wikipedia, user generated contents, knowledge bases, etc. Our goal is to find entity morphs corresponding to the given people as many as possible and ensure the found morphs are accurately referred to the corresponding people.

The challenges of identifying entity morphs include the following aspects:

1) It is hard to discover morphs through the co-occurrence of a morph and its target entity since people usually use either the morph or its target in an article rather than use both.
2) Morphs are always ambiguous, implicit and informal. So it is difficult to identify many morphs even for human if certain historical, cultural, or political knowledge is lacked.
3) Morphs evolve rapidly over time. So the methods of identifying entity morphs should cope with the dynamicity.

State of the Art: There are some prior methods for recognizing entity morphs or discovering name alias. Some methods [4, 5] take the known alias as seeds and find the patterns of the co-occurrence of the known alias and the real name, then find new alias based on the acquired patterns.

CIKM'17 , November 6–10, 2017, Singapore, Singapore
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4918-5/17/11...$15.00
https://doi.org/10.1145/3132847.3133123

However, such methods cannot identify alias if they never co-occur with the real name. Other methods [1, 3, 6] pre-define some rules to generate entity morphs. But they assume that candidate morph mentions are already discovered, which is actually a hard task in entity mention extraction since many morphs are informal terms. Besides, the generation rules are discrete, which can hardly cover all kinds of morphs.

Different from the current methods, our method doesn't need existing morphs as inputs. In order to imitate human's way of generating morphs for a specific person, we build the person-oriented knowledge graph to integrate fragmented knowledge together. Then morphs are generated according to the rules in the knowledge graph.

Contribution: This paper presents a Knowledge graph based methods to Identify Entity Morphs (KIEM for short) for a given person. The proposed method adopts the idea of both knowledge reasoning and statistic learning. First, a person-oriented knowledge graph is built. We define six kinds of relations in the knowledge graph. Then candidate morph mentions are generated according to the rules summarized from the knowledge graph. Finally, we propose to use a classification method to identify the target morphs from the candidate morph mentions. This paper involves the following contributions:

1) We propose a novel method to identify entity morphs based on knowledge graph.
2) We present a model of person-oriented knowledge graph and introduce the means of building the knowledge graph.
3) We conduct experiments on real-world dataset to validate the effectiveness of the proposed method.

2 METHODOLOGY

The proposed method compromises three main components: (1) building knowledge graph, (2) generating candidate morphs, and (3) identifying target morphs.

2.1 Building Knowledge Graph

The main objective of building a knowledge graph is to assemble comprehensive evidence of human's intentions and approaches to create entity morphs. By observing the existing morph mentions, we propose the following hypothesis.

Hypothesis 1: *The created morphs are inspired from the people, events, attributes, concepts and other entities which are related to the target entity.* For instance, the morph "苏牙 (Su-tooth)" is created to refer to the Uruguay soccer player "Luis Suarez" because of the event that Suarez bit an Italian player in a game of the 2014 World Cup.

Based on the above hypothesis, we utilize the knowledge graph technique to integrate the relations with the given person entity. The core node in the knowledge graph is the real name of the given person, other nodes are the related entities for generating morph mentions. Through investigating most existing morphs online, we define six kinds of relations involved in the knowledge graph.

Common Knowledge (CK): Given an entity name e, a relation of common knowledge is denoted as a 3-tuple $< e, r_{ck}, e_{ck} >$, where r_{ck} describes a certain relation on common knowledge such as "is-a", "birth-at", and "subclass-of". e_{ck} is the target entity in the relation.

The relations of common knowledge are used to describe the basic attributes and facts about the given person. Such relations are similar with the existing common knowledge bases (e.g. DBpedia, Yago, Freebase), which can be extracted from Wikipedia and similarly curated sources. The way people always use the common knowledge to create morphs is to switch the corresponding concept to another one with similar semantic meaning. For example, the existing morph "*Emperor Trump*" is derived from the following common knowledge relations:
$< Donald\,J.Trump, is\text{-}a, President >$,
$< President, subclass\text{-}of, ConutryLeader >$,
$< Emperor, subclass\text{-}of, ConutryLeader >$.

Related Events (RE). Given an entity name e, a relation of related events is denoted as a 3-tuple $< e, r_{re}, e_{re} >$, where r_{re} expresses that e participates in the event e_{re}. In this paper, we denote the event e_{re} as a set of representative keywords for simplicity.

The relations of related events can be extracted from news portals and social media. And the representative keywords can be extracted from the articles through some existing methods [7]. Take "范跑跑 (Fan Runner)" as an example, which is a famous morph referred to "范美忠 (Fan Meizhong)" in Chinese social media. This morph is created because Fan Meizhong, as a middle school teacher in Sichuan, ran by himself without helping his students when the famous "5.12 Wenchuan earthquake" happened. With the help of the relation of related events, our method can cope with the dynamicity of entity morphs, which is not considered in most of the existing methods.

Phonetic Similarity (PS). Given an entity name e, a relation of phonetic similarity is denoted as a 3-tuple $< e, r_{ps}, e_{ps} >$, where r_{ps} shows that e_{ps} is a character which is phonetically similar with a character in e.

Phonetic similarity makes the morphs sound similar to the real name but looks quite different. For example, "马饮酒(Ma Drinking)" is a morph of the former leader of Taiwan "马英九 (Ma Ying-jeou)". In Chinese, "马(Ma)饮(Yin)酒(Jiu)" and "马(Ma)英(Ying)九(Jiu)" have similar pronunciation. The relations to derive such morph are as follows:
$<$ 马英九, r_{ps}, 饮 $>$,
$<$ 马英九, r_{ps}, 酒 $>$.

The relations of phonetic similarity can be extracted from the Modern Chinese Pinyin Dictionary, which can automatically get all characters with the same pinyin for the characters in the real name.

Spelling Decomposition (SD). Given an entity name e, a relation of spelling decomposition is denoted as a 3-tuple $< e, r_{sd}, e_{sd} >$, where r_{sd} expresses that e_{sd} is the spelling decomposition of a character in e.

Since Chinese characters are ideograms, some characters can be decomposed into some basic units called radicals. Some of the radicals are also characters, thus we can use the radicals to replace the original character. For example, the famous historical figure "张飞(Zhang Fei)" can be morphed as "弓(Arrow)长(Long)飞(Fly)" through the relation of spelling decomposition $<$ 张飞 ,r_{sd}, 弓长 $>$.

Address Terms (AT). Given an entity name e, a relation of address terms is denoted as a 3-tuple $< e, r_{at}, e_{at} >$, where r_{at} expresses that e_{at} is an address term which can be used to call e.

Address terms are widely used to create morphs due to its simplicity. For example, "葛大爷 (Uncle Ge)" is a popular morph of a Chinses movie star "葛优 (Ge You)" in Chinese social media. This morph can be derived from the relation $<$ 葛优,r_{at}, 大爷 $>$.

Semantic Inference (SI). Given an entity name e, a relation of semantic inference is denoted as a 3-tuple $< e, r_{si}, e_{si} >$, where r_{si} expresses that e_{si} is semantically similar with part of e.

Semantic inference can generate morphs with semantic relation with the entity. The semantic relation can be extracted from some existing knowledge based such as WordNet [8] , HowNet [9]. For example, the relation $<$ 金日成,r_{si}, 太阳 $>$ helps to create the morph "金太阳 (Kim Sun)" for the former leader of North Korea "金日成 (Kim Il-sung)". This is because the character "日" in the real name has the meaning of "太阳 (Sun)".

2.2 Generating Candidate Morphs

In most of the existing methods, the morph generating rules are always independent from each other. This may cause that some morphs which use multiple rules cannot be recognized. For example, the morph "今太阳(Today Sun)" utilizes both the SI and the PS relations. However, one of the advantages of utilizing knowledge graph is to make fragmented knowledge work together. Thus, we develop a deductive method to generate candidate morphs by using the relations among the built knowledge graph.

Firstly, we define two kinds of nodes in the knowledge graph.

Definition 1 (NP): NP defines the target nodes in the relations CK, RE, AT, SI, which are used to combine with the characters in the real name (especially with the surname).

Definition 2 (NN): NN defines the target nodes in the relations PS and SD, which are used to replace the corresponding characters in the real name.

Based on the above definitions, the deductive method consists of two main steps:

1) Composing part of the real name with the NPs or replacing the real name totally with the NPs.

2) Substituting part of the mentions (both the real name and the mentions generated by the prior step) with NNs.

2.3 Identifying Target Morphs

Since the morph generation rules nearly enumerate all possible generation ways, the candidates generated might include some invalid or incorrect morphs. Thus, we must identify the target morphs which are correlated with the given person name. This problem of target morphs identification is modeled as a binary classification problem. That is, each pair of <candidate morph, real name> is classified to 'correlated' or 'uncorrelated'.

According to our observation of the usage of morph mentions, we propose the following hypothesis:

Hypothesis 2: *When people talk about a given person, the related entities or events will not change much no matter whether the morphs are used or not.* For instance, "*Black Mamba*" is a morph of the NBA star "*Kobe Bryant*". When people use "*Black Mamba*" to describe "*Kobe Bryant*", "*LA Lakers*", "*NBA games*", "*Shooting Guard*" are mostly mentioned in the context, which is almost the same when talking about "*Kobe Bryant*".

Our method firstly issue queries in a search engine with the given name and the candidate morphs as inputs respectively. Then we use the snippets returned by the search engines to construct the features of the given name and the candidates. According to hypothesis 2, related entities would co-occur in the features of the correct morphs and the real name. Thus, we adopts the entity topic model to form the features, which generates both word topics and entity topics by employing CorrLDA2 [10]. Then, we label some data and train a classifier by SVM. Finally, the classifier is used to decide whether a candidate morph is correlated.

3 EXPERIMENTS AND ANALYSIS

3.1 Setup

In this section, we present experiments to evaluate the quality of the output morphs of KIEM, in comparison with different methods.

Test Data. We collect 50 person entities from wikiptt[2] and the list of network languages in Mainland China[3]. These entities are involved in the category of politics, entertainment and sports and they all have quite many morphs on the web. Table 1 shows some example of the selected entities.

Table 1: Example of Selected Person Entities

Entity	Category	Sample Morphs
李登辉	politics	东瀛霸者，你等会
蒋介石	politics	蒋结实、老蒋、将校长
孙中山	politics	孙中三、孙大炮、国父
潘长江	entertainment	小陀螺、潘矮江
杨幂	entertainment	幂幂、嫩牛五方
姚明	sports	大姚、姚大个

[2] http://zh.pttpedia.wikia.com/wiki/Ptt

[3] http://zh.wikipedia.org/wiki/中国大陆网络语言列表

Methods under Comparison. In order to evaluate the performance of the proposed method in this paper, we compare our method KIEM with the two state-of-the-art methods. One is a lexical pattern-based method [4] and the other is a rule-based method [6].

Ground Truth. The method to decide whether the output morphs refer to the corresponding entities is based on artificial judgment. We ask 15 volunteers to judge the results of different methods. For each output morph, the volunteers firstly issue a query in Google with the morph. Then they check the first 10 pages of the query results. If they can recognize that the morph is definitely referred to the given entity through reading the returned snippets, then the morph is judged as a correct one.

Evaluation Metrics. To evaluate different methods for identifying entity morphs, we compare the outputs of different methods against the ground truth. The metrics for comparison are precision, recall and F-measure.

$$precision = \frac{|M_i \cap GT|}{|M_i|} \tag{1}$$

$$recall = \frac{|M_i \cap GT|}{|GT|} \tag{2}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

where M_i is the set of output morphs by the *i*-th method, GT is the set of correct morphs mentioned in ground truth statements.

3.2 Results Analysis

Table 2 shows the comparison results on precision, recall and F-measure. From the results on all categories, we can observe that KIEM outperforms the other two both on precision and recall. The patterned-based performs better than the rule-based method on precision but worse on recall. In our consideration, the reason is that the rule-based method generates too many invalid morphs while the pattern-based method can only identify the morphs that co-occur with the real names. From the results on each category, we can find that KIEM and rule-based both perform best in the category of politics. This is because the morphs for politicians are always based on some rules that can help them evade the censor ship. However, the morphs for entertainment stars or sports stars are always created by their fans and usually co-occur with the real names. Thus makes the pattern-based method perform better on the categories of sports and entertainment. However, KIEM also performs better than the pattern-based method on the categories of entertainment and sports. This is because KIEM constructs relations between the real name and related entities, which would recognize more forms of the co-occurrence of the real name and morphs than the pattern-based method.

Table 2: Comparison Results

Categories	Metric	KIEM	Pattern-based	Rule-based
Overall	P	**0.81**	0.49	0.23
	R	**0.68**	0.14	0.43
	F	**0.74**	0.22	0.30
Politics	P	**0.82**	0.47	0.24
	R	**0.69**	0.12	0.44
	F	**0.75**	0.19	0.31
Entertainment	P	**0.68**	0.52	0.14
	R	**0.58**	0.35	0.40
	F	**0.62**	0.42	0.21
Sports	P	**0.65**	0.61	0.24
	R	**0.40**	0.34	0.37
	F	**0.50**	0.44	0.29

4 CONCLUSIONS

In this paper, we have introduced a knowledge graph based method which targets at solving the problem of identifying entity morphs. By deeply analyzing the common rules of creating morphs and the unique challenges of this problem, we build a person-oriented knowledge graph to integrate fragmented knowledge on creating morph mentions and use the knowledge to generate candidate morphs. Then we propose a classification method to achieve the target morphs from the candidates. Finally, extensive comparative experiments demonstrate the efficiency of our method.

In the future, we will refine our method of generating candidate morphs by using some heuristic algorithms which can avoid generating too many useless candidates. Besides, we will extend our method to identify morphs of other subjects such as events and organizations.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China under Grant No. 61702500.

REFERENCES

[1] H. Huang, Z. Wen, D. Yu, H. Ji, Y. Sun, J. Han, and H. Li. 2013. Resolving Entity Morphs in Censored Data. *Meeting of the Association for Computational Linguistics.* (Aug. 2013). 1083-1093.

[2] L. Chen, C. Zhang, and C. Wilson. 2013. Tweeting under pressure: analyzing trending topics and evolving word choice on sina weibo. *ACM Conference on Online Social Networks.* (Oct. 2013). 89-100.

[3] B. Zhang, H. Huang, X. Pan, S. Li, C. Y. Lin, H. Ji, K. Knight, Z. Wen, Y. Sun, and J. Han. 2015. Context-aware Entity Morph Decoding. *Meeting of the Association for Computational Linguistics.* (Aug. 2015). 586-595.

[4] D. Bollegala, Y. Matsuo, and M. Ishizuka. 2011. Automatic Discovery of Personal Name Aliases from the Web. *IEEE Transactions on Knowledge & Data Engineering*, 2011, 23 (6): 831-844.

[5] D. Bollegala, T. Honma, Y. Matsuo, and M. Ishizuka. 2008. Mining for personal name aliases on the web. *International Conference on World Wide Web.* (April 2008). 1107-1108.

[6] B. Zhang, H. Huang, X. Pan, H. Ji, K. Knight, Z. Wen, Y. Sun, J. Han, and B. Yener. 2014. Be Appropriate and Funny: Automatic Entity Morph Encoding. *Meeting of the Association for Computational Linguistics.* (Aug. 2014). 706-711.

[7] K. S. Dave, and V. Varma. 2010. Pattern based keyword extraction for contextual advertising. *ACM International Conference on Information and Knowledge Management.* (Oct. 2010). 1885-1888.

[8] C. Fellbaum, and G. Miller. 1998. WordNet:An Electronic Lexical Database. *MIT Press.* 1998.

[9] Z. Dong, Q. Dong, and C. Hao. 2010. HowNet and its computation of meaning. *International Conference on Computational Linguistics: Demonstrations.* (Aug. 2010). 53-56.

[10] Hou, L., Li, J., Wang, Z., Tang, J., Zhang, P., and Yang, R. (2015). Newsminer: multifaceted news analysis for event search. *Knowledge-Based Systems*, 2015, 76: 17-29.