A Unified Approach for Short Question Entity Discovery and Linking

Qin Wei[1]，Jiong Zhang[2]，Huimin Zhang[2]

[1] University of Shanghai for Science and Technology, Shanghai 200093, China
[2] Shanghai Putao Technology Co., Ltd.,Shanghai 200233, China
{zhanghuimin,zhangjiong,weiqin}@putao.com
157730675@st.usst.edu.cn

Abstract

The problem of entity discovery and linking(EDL) in short questions aims to find the entities the questions focused on and disambiguate them, usually over linked-data sources. For Chinese questions, the problem mainly involves three tasks: the segmentation of words and phrases in questions, the segment disambiguation and the mapping of mentions to semantic entities. In this paper, we propose an integer linear program (ILP) based method to solve the three tasks jointly. Our solution harnesses the rich feature types provided by the question context and the linked data source, CN-DBpedia in the experiment, to constrain our semantic-coherence objective function and a genetic algorithm(GA) is used to tune the parameters. In the evaluation of CCKS2017 shared task one, our approach achieves a f1 score of 0.804 in the mention discovery and 0.56 in the entity linking, and ranks 1st among all the 17 teams according to the f1 score of entity linking.

1 Introduction

Entity Discovery and Linking (EDL) in Natural Language Processing (NLP) is the task of matching entity mentions in texts to a unique identifier in linked-data sources, such as CN-DBpedia, and it is becoming a hot topic as the linked data grows. Unlike conventional tasks of Named Entity Recognition (NER), which focus on identifying the occurrence of an entity and its type, but not the specific unique entity that the mention refers to, EDL makes a further step in understanding texts, thus playing a critical role in the construction of the upper applications, such as the Information Retrieval (IR) and Knowledge Based Question Answering (KBQA) systems.

EDL is commonly divided into two sub tasks: Mention Detection (MD) and Entity Linking (EL). MD is concerned with identifying potential mentions of entities in the text and EL involves mapping mentions to semantic entities. EDL is complex and challenging due not only to the ambiguity of word and phrase senses but also entity mentions, which are affected by the context of words and phrases, the similarity of mentions and entities, the prior of entities, the coherence among entities and etc.

Approaches have been proposed in literature to solve EDL. The majorities treat the two tasks in EDL separately, which can be distinguished into two main types: rule-based approaches and machine learning models. Rule-based approaches make good efforts in linguistic analysis [1] [2] [3] and can build practical systems in limited

periods, thus getting good performances in the early related shared tasks. Machine learning models such as Maximum Entropy (ME) [4], generative models [5], ranking methods[6] and etc., benefitting from the data explosion in the last decade, good at balancing the precision and recall, is becoming more and more dominate. Suffering from cascade errors, gaps to the theoretical best performance exist for these separate approaches. Some jointly approaches are also reported [7] [8], which are good at taking the affecting factors of EDL to a single model but lack methods for model parameter tuning.

This paper presents our approach for CCKS2017 shared task, Question Entity Discovery and Linking (QEDL) in Chinese. One more sub task is raised, the segmentation of words and phrases in questions, compared to EDL, due to the boundary lack of Chinese words. The three sub tasks are jointly solved by an Integer Linear Program (ILP) based model tuning parameter by the Genetic Algorithm (GA). An f1 score of 0.804 in the mention discovery and 0.56 in the entity linking have been achieved in the evaluation.

The paper is structured as follows. After describing the four steps of the online predicting framework in section 2, we discuss the joint disambiguation step in detail in section 3. Section 4 presents the offline parameter tuning of the online model. The evaluation results are outlined in section 5. Finally, we review the main conclusions and preview the future work.

2 Framework

As shown in figure 1, given a Chinese short question, our online approach takes four steps for QEDL: word detection, mention discovery, entity mapping and joint disambiguation. A question sentence is processed as a sequence of characters, qNL = (t_0, t_1,… t_n) while a word is a contiguous subsequence of the character sequence, w_{ij} = (t_i, t_{i+1}, … t_j) , 0<=i<=j<=n. The input question is handled by the pipeline in figure 1.

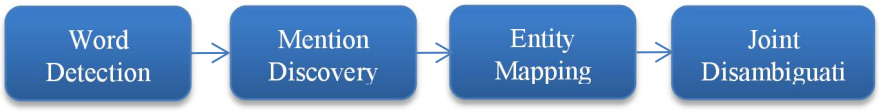

Figure 1: Architecture of the online system

2.1 Word Detection

Words are detected by “jieba”, a common used Chinese text segmentation. In the cut for search mode of “jieba”, all the word candidates are generated and put to a word set. For sample question “李娜是在哪一年拿的澳网冠军”, the word set contains the following candidates: “李娜”,“是”,“在”,“哪一年”,“一年”,“拿”,“的”,“澳网” and “冠军”。

2.2 Mention Discovery

Mentions are discovered using CN-DBpedia, by querying the contiguous subsequences of the questions. Subsequences are made mention candidates, which will be added to the word set containing all the word candidates, by the existence of the query results. For the sample question, the mention candidates, “李”，“李娜”，“一”，“一年” and etc., are added to the word set, with duplication removed.

2.3 Entity Mapping

A mention candidate can be mapped to multiple semantic entities. During this step, a semantic entity mapping space for the mentions is constructed. “李娜” in the above sample question is mapped to “李娜（中国女子网球名将）”，“李娜（南京师范大学讲师）”，“李娜（2016年陈可辛导演电影）” and etc., in the mapping space.

2.4 Joint Disambiguation

During this step, the word boundary determination, mention disambiguation and semantic entity disambiguation are solved jointly by calculating a disambiguation graph. For the sample question, by decoding the outcome graph, we can get entity mentions as “李娜” and “澳网” and entity linking as “李娜（中国女子网球名将）” and “澳大利亚网球公开赛”. The details of this step is described in section 3.

3 Joint Disambiguation

As the disambiguation of one word, mention and entity can influence the others, a disambiguation graph encoding all possible mappings is constructed. To simplify the problem, we model the problem as an ILP, rather than graph models.

3.1 Overlap-Word-Entity Graph

A weighted, undirected Graph DG= (V, E) is defined with words Vw, entities Ve and the word overlap constrains Vo as nodes. The graph for the sample question is shown in figure 2, from which, we can see the edges, in which the word-entity edges tie closely to the final predicting results, are indicated by solid and dashed lines corresponding to 1 and 0 are assigned to variables in ILP.

Features and Constraints are expressed by the weighted edges in the Graph and by optimizing the final objective function, the best word and entity nodes are selected. The features and constraints harnessed in the model are presented below.

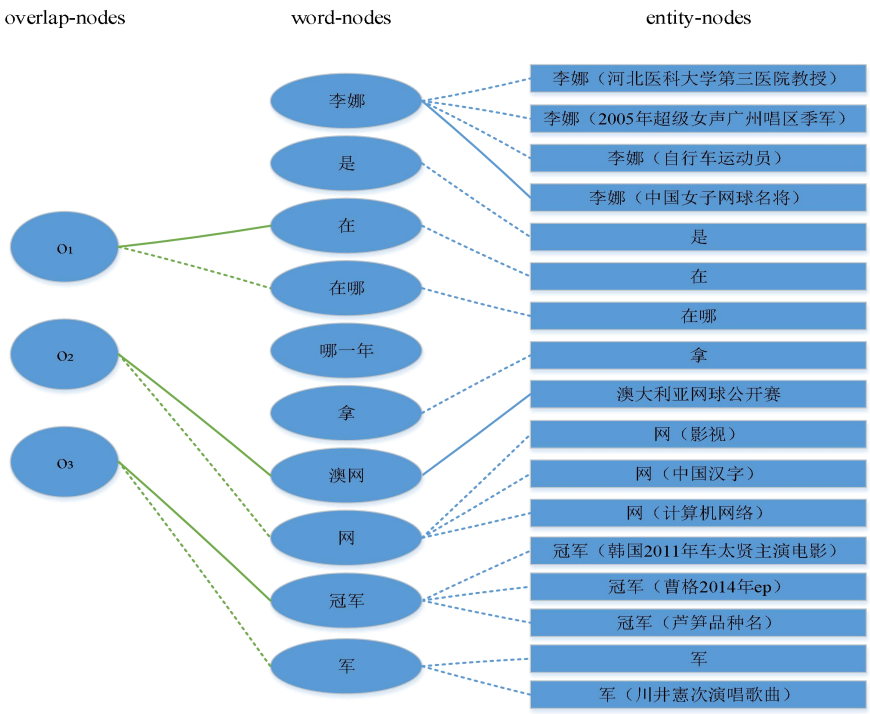

Figure 2: Overlap-Mention-Entity Graph Example

3.2 Features and Constraints

Mention Prior (Prior(mention)): mention prior are related to the mention character length, Part Of Speech (POS) of the duplicated word, whether it is covered by other words and some linguistic features such as numbers, special characters are contained.
Entity Prior (Prior(entity)): entities have different priors and the number of attributes, special attributes, special attribute values are chosen to signify the prior of entities in CN-DBpedia.
Entity Context (Sim(entity-context)): some context words indicate higher priorities of entities which is captured by entity context similarity.
Mention to Entity Similarity(Sim(M-E)): this feature captures the mention and entity name matching degree.
Entity Coherence(Sim(coh)): some entities are high correlated, which usually are the question topic related entities, by sharing the same category, quoting each other, related to the same other entity and etc., expressed by the entity coherence similarity.
Overlap Constraints (Con (overlap)): unique character cannot exists in two words.
Mention Entity Constraints (Con (mention-entity)): one mention corresponds to one entity at most.
Mention Number Constraints (Con (mention-number)): short question usually concerns one topic and the related mentions are limited, so we added mention number constrain, which penalize more as the number of mentions increases.

3.3 Over-all Objective Function

Our framework combines mention prior, entity prior, entity-context prior, similarity between entity and mention, and coherences between entities into a combined objective function.

$$\sum_{i=1}^{k_1}\sum_{t=1}^{p_1}\alpha_t * \text{prior}(m_i) + \sum_{i=1}^{k_2}\sum_{j=1}^{k_i}\sum_{t=1}^{p_2}\beta_t * \text{prior}(e_{ji}) + \sum_{i=1}^{k_2}\sum_{j=1}^{k_i}\sum_{t=1}^{p_3}\gamma_t * \text{sim}(e_{ji}, \text{cxt}(m_i)) + \sum_{i=1}^{k_2}\sum_{j=1}^{k_i}\sum_{t=1}^{p_4}\mu_t * \text{sim}(m_i, e_{ji}) + \sum_{i=1}^{k_2}\sum_{j=1}^{k_i}\sum_{n=i+1}^{k_2}\sum_{m=1}^{k_n}\sum_{t=1}^{p_5}\lambda_t * \text{sim}(e_{mn}, e_{ji}) - \sum_{t=2}^{4} c_t * (\text{sum}(m_i) - t + 1)) = \max! \quad (1)$$

Subject to:

$$\sum_{i=1}^{k} \text{I}(m_i, j) \le 1, \quad \text{I}(m_i, j) = \begin{cases} 1, & \text{if } j \in \text{scope}(m_i) \\ 0, & \text{other} \end{cases} \quad (2)$$

$$0 \le \text{I}(e_i) \le \text{I}(m_i) \le 1, \quad \text{I}(e_i), \text{I}(m_i) = \begin{cases} 1, & \text{if } e_i \text{ or } m_i \text{ exists} \\ 0, & \text{other} \end{cases}$$

Where $\sum_{t=1}^{p1}(\alpha_t)^2 + \sum_{t=1}^{p2}(\beta_t)^2 + \sum_{t=1}^{p3}(\gamma_t)^2 + \sum_{t=1}^{p4}(\mu_t)^2 + \sum_{t=1}^{p5}(\lambda_t)^2 + \sum_{t=2}^{4}(c_t)^2 = 1$, $\text{cxt}(m_i)$ denotes the context of mentions, $\text{sum}(m_i)$ represents the number of all the mentions, do penalize when the number of mentions is over 2, 3 or 4, each feature correspond to one coefficient, which changing by the GA tuning. $\text{scope}(m_i)$ is the start position and end position interval of m_i in the question sentence. Section 3.2 gives details on each of these components.

3.4 ILP Processing

We use binary variables to indicate whether the nodes and edges are selected and integrate the features and constraints to the ILP objective function and constraints as shown in equation 1 and make the objective function linear with introducing some new variables and the spinoff constraints. It seems a sophisticated ILP, but for the questions are short, it is within the regime of modern ILP solvers. In Our Experiment, we use Pulp and achieved run-times, usually less than one second.

4 Parameter Tuning

The parameters in the objective function in ILP, about 30 in quantity, are optimized by GA, a random search and optimization method based on natural selection and genetic mechanism of the living beings [10], without calculating the gradients. As the target parameters are floats, real number coding method are elected.

The CCKS2017 shared task one use f1 score in mention discovery and entity linking, We use score in equation (5) as our fitness function in GA, which grows bigger as the f1 scores increase.

$$score\ \_e = |A \cap B| / (alpha * |A - B| + |B|) * beta \quad (3)$$

$$score\ \_l = |A \cap B| / (alpha * entityDiff\ (A, B) + |B|) * beta \quad (4)$$

$$score = gamma * score\ \_e + (1. - gamma) * score\ \_l \quad (5)$$

Where, $alpha$ is the recognition rate adjustment coefficient, $beta$ is the bonus coefficient, $gamma$ is the fitting coefficient and $entityDiff(A, B)$ is the count of different elements in set A and B .

5 Evaluation

We evaluate our system in CCKS2017 shared task one, QEDL. 1400 questions are provided as training data, with mentions and entities annotated while another 750 questions as test data without annotated information. The training and testing procedures are carried out on CN-DBpedia, which consists of 16,601,597 Baike entities and 213,945,421 Baike relationships.

Table 1: Output Examples of Our System.

question	mention	entity
李娜是在哪一年拿的澳网冠军？	李娜|||澳网	李娜（中国女子网球名将）|||澳大利亚网球公开赛
《天天向上》门票多少钱一张？	天天向上	天天向上（湖南卫视娱乐节目）
PPS和风行看电影哪个好?	PPS|||风行	pps（网络电视软件）|||风行（网络视频软件）
日本拍哪些中日战争电影	日本|||中日战争	日本|||中日战争（中日战争）
关于曹操、关羽、刘备、诸葛亮的介绍	曹操|||关羽|||刘备|||诸葛亮	曹操|||关羽|||刘备|||诸葛亮（三国时期蜀汉丞相）

As Shown in Table 1, our model makes reasonable predictions in the QEDL. Further Experiments show the f1 score of 0.804 in the mention discovery and 0.56 in the entity linking are achieved on the test data, and over 0.10 are obtained on both scores over the second team in the shared task.

6 Conclusion and Future Work

In this paper, an ILP based approach has been proposed for CCKS2017 shared task one. The approach harnesses the rich feature types provided by the question context and the linked data source to constrain the semantic-coherence objective function using a GA to tune the parameters, achieving the best performance in the task. Future work includes considering additional feature mining, improving online model calculating efficiency and its universality in other corpus.

References

1. Soraluze, Ander, et al. "Mention detection: First steps in the development of a Basque coreference resolution system." KONVENS. (2012).
2. Zhekova, Desislava, and Sandra Kübler. "UBIU: A language-independent system for coreference resolution." Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics.(2010).
3. Lee, Heeyoung, et al. "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task." Proceedings of the fifteenth conference on computational natural language learning: Shared task. Association for Computational Linguistics.(2011).
4. Sil A, Florian R. The IBM systems for English entity discovery and linking and Spanish entity linking at TAC 2014[C]//Text Analysis Conference (TAC), Gaithersburg, Maryland, USA. (2014).
5. Han X, Sun L. A generative entity-mention model for linking entities with knowledge base[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 945-954.(2011).
6. Shen W, Wang J, Han J. Entity linking with a knowledge base: Issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 27(2): 443-460.(2015),
7. Yahya, Mohamed, et al. "Natural language questions for the web of data." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, (2012).
8. Hoffart, Johannes, et al. "Robust disambiguation of named entities in text." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, (2011).
9. Lingen Ji. :Survey On Genetic Algorithm. Computer Applications and Software, 21(2), 69-73.(2004).
10. Jike Ge, Yuhui Qiu, Chunming Wu, &Guolin Pu.Survey On Genetic Algorithm. Computer Application Research, 25(10), 2911-2916. (2008).