FALCON: An Entity and Relation Linking Framework over DBpedia

Ahmad Sakor[1], Kuldeep Singh[2], and Maria-Esther Vidal[3,1]

[1] L3S Center Germany, `sakor@l3s.de`
[2] Nuance Communications, Germany, `kuldeep.singh1@nuance.com`
[3] TIB, Hannover, `maria.vidal@tib.eu`

Abstract. We tackle the problem of entity and relation linking and present FALCON, a rule-based tool able to accurately map entities and relations in short texts to resources in a knowledge graph. FALCON resorts to fundamental principles of the English morphology (e.g., compounding and headword identification) and performs joint entity and relation linking against a short text. We demonstrate the benefits of the rule-based approach implemented in FALCON on short texts composed of various types of entities. The attendees will observe the behavior of FALCON on the observed limitations of Entity Linking (EL) and Relation Linking (RL) tools. The demo is available at https://labs.tib.eu/falcon/.

Keywords: NLP · Entity Linking · Relation Linking · Background Knowledge · English morphology

1 Introduction

In the era of digitisation, data availability has exponentially grown in the last years and a similar growth rate is expected in the next decade. Although a large volume of data is presented in structured formats, e.g., relational tables, a large number of attributes are associated with values in the form of short text. Named Entity Recognition and Disambiguation (NER and NED), relation linking (RL) tasks annotate surface forms in text with the corresponding reference mentions in a knowledge base such as Wikipedia. With the emergence of Knowledge Graphs (KGs) that represent Wikipedia data in a structured and semantic format such as DBpedia [1], and Wikidata [5], retrieval-based applications such as question answering (QA) or keyword-based semantic search systems are empowered to provide more cognitive capabilities. There are existing approaches which address EL and RL tasks either jointly or independently. However, these approaches mostly fail in case of a short text (e.g., question or keywords based query) because a short text does not provide sufficient context which is essential for the disambiguation process. Several efforts have been made to develop novel approaches for relation extraction but limited literature is available for approaches targeting linking of extracted relations to its corresponding KG. This
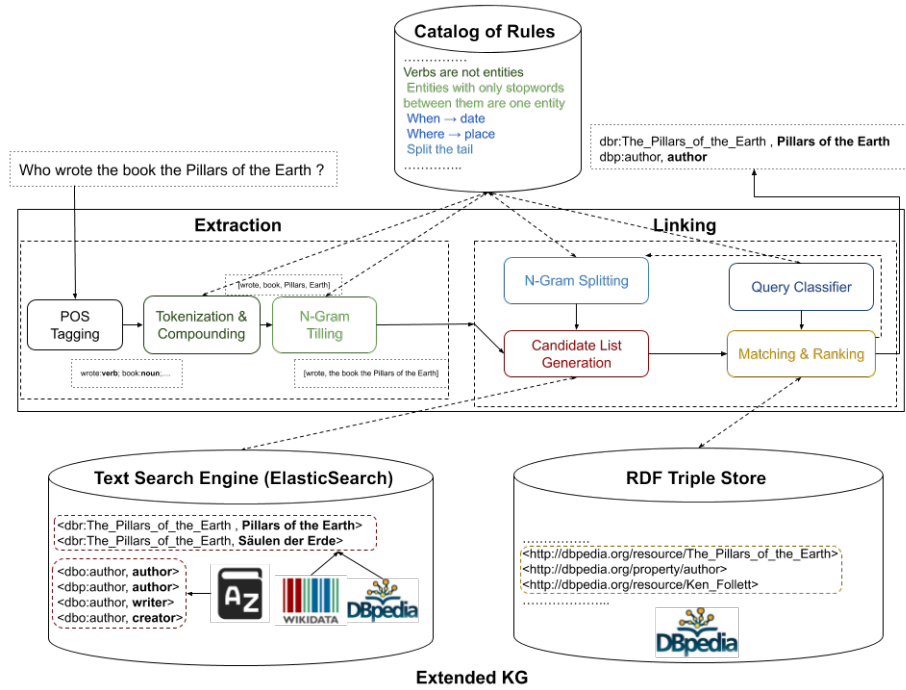
Fig. 1. The FALCON architecture. FALCON consists of two building blocks: 1) An extended knowledge graph which is built by merging information from various knowledge sources such as DBpedia, Wikidata, Oxford Dictionary. 2) FALCON architecture that has several modules focusing on surface form extraction and linking them to KG.

paper presents the demonstration of the FALCON approach [3], which jointly attempts entity and relation linking of short text to the mentions in DBpedia.

2   The FALCON architecture

The FALCON architecture is depicted in Figure 1. FALCON receives as input short texts and outputs a set of entities and relations extracted from the text; each entity and relation in the output is associated with a resource in a knowledge graph. FALCON resorts to an extended knowledge graph and a catalog of rules for performing entity and relation linking. The extended knowledge graph integrates information from various knowledge sources, e.g., DBpedia, Wikidata, Oxford Dictionary, and Wordnet. Additionally, it comprises alignments between nouns and entities in these knowledge sources. Alignments are stored in a text search engine, e.g., ElasticSearch, while the knowledge sources are maintained in an RDF triple store accessible via SPARQL endpoints. The rules that represent the English morphology are maintained in a catalog; a forward chaining inference process is performed on top of the catalog during the tasks of extraction

and linking. FALCON also comprises several modules which identify and link entities and relations to DBpedia knowledge graph. These modules implement POS Tagging, Tokenization & Compounding, N-Gram Tilling, Candidate List Generation, Matching & Ranking, Query Classifier, and N-Gram Splitting.

3   Demonstration of Use Cases

We motivate our work by demonstrating how FALCON performs over short texts composed of a great variety of entities and relations; these use cases represent challenges to the state-of-the-art tools for entity and relation linking.

Effect of Capitalization on EL tools. Consider the question 'When was University of Edinburgh founded', where the entity `University of Edinburgh` has one word (i.e., 'of') starting with lowercase letters. FALCON correctly identifies and links this entity to `dbr:University_of_Edinburgh`–corresponding DBpedia resource. Existing best performing approaches (i.e., TagMe and DBpedia-spotlight) assume that every entity starts with a capital letter; in consequence, entities in this type of short texts cannot be recognized by these tools. [3].

Effect of Implicit/Explicit Entities on EL tools. The vocabulary mismatch problem is common for text paraphrasing and significantly affects the behavior of EL tools [4]. TagMe and DBpedia spotlight can correctly link the entity in the question 'How high is Colombo Lighthouse?', but they fail if the question is rephrased to 'How high is the lighthouse in Colombo?' due to the vocabulary mismatch problem. In the first question, the entity `Colombo Lighthouse` matches to the DBpedia entity `dbr:Colombo_Lighthouse` which is not the case in the rephrased question where the entity is `lighthouse in Colombo`). We demonstrate how FALCON links correctly these entities in both cases.

Effect of the Number of Words in an Entity Label on EL tools. During question answering, an entity with more than three words negatively affects the linking performance of EL tools [3]. Both Tagme and DBpedia-Spotlight are not able to link the entity present from the question 'Who wrote the book The Pillars of the Earth?' [4] where the entity label ('The Pillars of the Earth') has five words; we demonstrate how FALCON links this entity correctly.

Effect of Ambiguity of Question on RL tools. EARL[5] and Rematch [2] are top performing RL tools for question answering over two different datasets QALD-5 and LC-QuAD, respectively. For the question 'When did princess Diana die', Rematch correctly recognizes the relation `die` and links it to `dbo:deathYear`. However, when the question is slightly changed to "Where did princess Diana die?" in which the expected relation is `dbo:deathPlace`, both tools fail to understand the ambiguity of the question and cannot identify the correct DBpedia resources. FALCON recognizes the slight difference in both questions, and overcomes the ambiguity by considering the question head word (When/Where).

Effect of Hidden Relation in a Question on RL tools. Questions are typically relatively short and sometimes there is no natural language label for

---

[4] A question from LC-QuAD dataset

[5] http://sda.cs.uni-bonn.de/projects/earl/

the relation. For example, the QALD-6 question 'Who is starring in Spanish movies produced by Benicio del Toro?' where one of the expected relations is `dbo:country` for which no relation label is present. For the previous question, EARL and ReMatch cannot identify these hidden relations. While FALCON identifies and links these hidden relations correctly.

Effect of Derived Word Form of Relation Label on RL tools. Considering the question 'Was Ganymede discovered by Galileo Galilei?' in which the relation label `discovered` should be linked to the DBpedia entity `dbo:discoverer`. The word `discoverer` is the derived word form of relation label `discovered`, and due to this, both tools fail to provide correct relation linking. While FALCON can link this entity correctly during the matching and ranking step.

4   Conclusions

The FALCON rule-based tool implements two novel concepts. First, a fused knowledge graph comprising several complimentary semantic and linguistic resources which are employed as background knowledge. Second, the representation of the English morphology using rules which allows for effectively identifying entities and relations in short texts in English. Furthermore, alignments between nouns and entities in the extended background knowledge graph enable effective EL and RL. As a proof of concept, we will demonstrate FALCON over DBpedia, i.e., the alignments are set to relate nouns to DBpedia. However, there is no specific assumption in the FALCON approach about the structure or schema of the underlying knowledge graphs. Thus, FALCON could equally perform if the alignments would be defined over any other knowledge graph.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In: ISWC. pp. 722–735. Springer (2007)
2. Mulang', I.O., Singh, K., Orlandi, F.: Matching natural language relations to knowledge graph properties for question answering. In: Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017. pp. 89–96 (2017)
3. Sakor, A., Mulang, I.O., Singh, K., Shekarpour, S., Vidal, M.E., Lehmann, J., Auer, S.: Old is gold: linguistic driven approach for entity and relation linking of short text. In: Proceedings of the 2019 NAACL HLT. pp. 2336–2346 (2019)
4. Singh, K., Lytra, I., Radhakrishna, A.S., Shekarpour, S., Vidal, M., Lehmann, J.: No one is perfect: Analysing the performance of question answering components over the dbpedia knowledge graph. CoRR **abs/1809.10044** (2018)
5. Vrandecic, D.: Wikidata: a new platform for collaborative data collection. In: Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume). pp. 1063–1064. ACM (2012)