Entity Linking for Open Information Extraction

Arnab Dutta and Michael Schuhmacher

Research Group Data and Web Science, University of Mannheim, Germany
{arnab,michael}@informatik.uni-mannheim.de

Abstract. Open domain information extraction (OIE) projects like Nell or ReVerb are often impaired by a schema-poor structure. This severely limits their application domain in spite of having web-scale coverage. In this work we try to disambiguate an OIE fact by referring its terms to unique instances from a structured knowledge base, DBpedia in our case. We propose a method which exploits the frequency information and the semantic relatedness of all probable candidate pairs. We show that our combined linking method outperforms a strong baseline.

1 Introduction

In the recent past, there have been major developments in the area of open domain information extraction (OIE). Projects like Nell [2] or ReVerb [5] have introduced an era of *open* information extraction systems which are characterized by their web-scale coverage at the expense of a poor schema. On the other extreme, Wikipedia based extraction systems like DBpedia [1] or Yago [11] provide more structured information but have limited coverage.

The data maintained by OIE systems is important for analyzing, reasoning about, and discovering novel facts on the web and has the potential to result in a new generation of web search engines [4]. But the lack of a proper schema severely limits the applicability of such data. Moreover, facts from the OIE are often too ambiguous. For instance, a typical fact extracted by Nell might be *bookwriter*(*imperialism, lenin*). While we might have an intuitive understanding of the property *bookwriter*, it is difficult to determine the correct references of the subject and object terms within the triple. Here, the surface form object *lenin* can refer to Vladimir Lenin (the Russian political theorist), Lenin (a nuclear icebreaker) or the Lenin Prize. For that reason, we need to disambiguate the Nell terms to uniquely identifiable instances. We opt to use DBpedia as the structured knowledge-base providing globally unique URIs for each instance.

In general, the entity linking task is to match surface form mentions from a natural language text to the corresponding knowledge base instances. However, in this work, we focus on linking ambiguous Nell subjects and objects to DBpedia. While established Entity Linking systems, like e.g. DBpedia Spotlight [8] or Aida [7], exploit, besides other features, mainly the context of the entity within the text, in our case of OIE triple linking, this context information is in many cases not available or does not exist. This sets our problem setting apart from the traditional linking task.

E. Métais, M. Roche, and M. Teisseire (Eds.): NLDB 2014, LNCS 8455, pp. 75–80, 2014.

2 Entity Linking Methods

In the following, we present first a simple, yet very strong, baseline method, namely the Frequency-based Entity Linking as used in [3]. Second, we introduce a knowledge-based approach which exploits the DBpedia ontology itself, following the DBpedia graph exploration method introduced by [10]. Last, we propose a Combined Entity Linking approach, which incorporates the frequency-based with the graph-based approach. Note that in this work, we do not focus on mapping predicates from Nell to corresponding DBpedia properties because (a) existence of corresponding DBpedia properties cannot be guaranteed and (b) an exact analogous mapping may not be possible, for instance the Nell property *agentcollaborateswithagent*, could be mapped to any one of `dbp:influences`, `dbo:publisher` or `dbo:employer`.

Frequency-Based Entity Linking. A simple, yet high performing approach for mapping a given surface form (Nell subject/objects in our case), to its corresponding DBpedia (or Wikipedia) instance is to link to its most frequent candidate entity [9]. Even though this approach does not take any context information into account, it has been proven to be effective not only for text entity linking, but also for Nell triple linking [3]. We thus use it as a baseline method. For obtaining the frequencies, the intra-Wikipedia links connecting anchor texts to article pages were exploited (using WikiPrep [6]). The assumption was that, the maximum number of outgoing links from an anchor to a particular article marks the article as the most probable entity referred to by its surface form (the anchor). Formally, if an anchor e refers to N Wikipedia articles $A_1, \ldots, A_N$ with $n_1, \ldots, n_N$ respective link counts, then the conditional probability P of e referring to A_j is given by, $P(A_j|e) = n_j / \sum_{i=1}^{N} n_i$. Thus, the pair (e, A_j), henceforth called subject/object-instance-mapping, is awarded the probability P. For every Nell triple, we can use this approach since each DBpedia entity is equal to its Wikipedia article title. We rank the candidates on descending P and define a *top* ranked list as $E_{Subj|top-k}$ (for subject mappings) and $E_{Obj|top-k}$ (for object mappings). Since every mapping of a subject is independent of the object mapping, we compute the prior probability as $P_{prior} = P_{Subj} P_{Obj}$. We select the DBpedia subject-object pair with the highest prior probability.

Graph-Based Entity Linking. We assume that the subject and object connected by a Nell predicate are related to each other and that some relationship can be found within the DBpedia knowledge base. This motivates us to exploit the latent contextual connectivity between Nell terms instead of relying just on the most frequent entity.

We obtain the likelihood of each possible pair of subject-object candidates by computing the semantic relatedness [12] between subject and object. As we do not want to make any assumptions about the existence or the type of the DBpedia properties to be taken into account, we adapt the property-agnostic approach presented in [10] and summarized as below.

1) We consider all combinations $E_{Subj|top-5} \times E_{Obj|top-5}$ and compute each pairwise cheapest path, treating DBpedia as a semantic network (see [10] for details).
2) We weigh the DBpedia graph edges, thus automatically capturing the importance of different property edges, by an information-content- based measure (CombIC) which was reported as the best of the graph-weighting schema proposed by [10] for computing semantic similarity.
3) We select the subject-object pair from $E_{Subj} \times E_{Obj}$ which has the minimum path cost, on the weighted graph. The path cost between two entities is calculated as the sum of the edge costs along their undirected connecting path and is normalized as probabilities to P_{graph}

As result, we jointly disambiguate subject and object to their most semantically similar DBpedia candidate entities.

Combined Entity Linking. Our last approach is motivated by the fact that both approaches the frequency-based and the graph-based linking have an individual weakness, but can complement each other. The former exploits the empirically obtained frequency data about common surface-form-to-instance mappings, however, it cannot incorporated the information that subject and object should most likely be related somehow. But this information is used by the graph-based linking, which finds this vague relationship between subject and object in the background knowledge base DBpedia, however, ignoring the important frequency information. Consequently, we opt for a linear combination of the two approaches and select the subject-object combination with the highest combined probability

$$P_{comb} = \lambda P_{graph} + (1 - \lambda) P_{prior}$$

where the weighting factor λ is set to 0.5 initially, thus giving equal influence to the graph and the frequency information. With this combination, we give preference to those subject-object combinations, having individually high likelihoods and which are also closely semantically related in the DBpedia knowledge-base.

3 Experiments

Dataset and Metric. We use the gold standard from [3][1], which consists if 12 different Nell properties with 100 triples each that have been manually linked to their correct DBpedia entities. For our evaluation, we excluded the predicate *companyalsoknownas*, as it contains actually not distinct subjects and object, but only different surfaces forms for the same entity, e.g. *companyalsoknownas(General Motors, GM)*. As metric, we use Precision (P), Recall (R) and F-measure (F_1), and evaluate each subject and object mapping individually, thus also accounting for partially correct triple.

[1] Downloaded from `https://madata.bib.uni-mannheim.de/65/`

Table 1. Performance scores of our proposed methods and the baseline. Best F_1 values for each predicate is marked in bold.

	Frequency-based			Graph-based			Combined		
	P	R	F_1	P	R	F_1	P	R	F_1
actorstarredinmovie	80.7	82.0	81.3	89.8	91.2	90.5	91.4	92.8	**92.1**
agentcollaborateswithagent	81.6	85.9	**83.7**	69.3	72.9	71.1	81.6	85.9	**83.7**
animalistypeofanimal	85.7	88.0	**86.8**	62.4	64.1	63.3	85.2	87.5	86.3
athleteledsportsteam	88.6	85.5	87.0	87.0	84.0	85.5	91.7	88.5	**90.1**
bankbankincountry	81.7	77.6	**79.6**	68.3	64.8	66.5	81.7	77.6	**79.6**
citylocatedinstate	79.0	79.4	79.2	81.5	81.9	81.7	86.0	86.4	**86.2**
bookwriter	82.2	83.1	82.6	83.8	84.7	84.2	87.6	88.5	**88.0**
personleadsorganization	83.6	79.0	81.2	78.4	74.0	76.1	84.8	80.1	**82.4**
teamplaysagainstteam	81.8	81.8	81.8	61.0	61.0	61.0	85.6	85.6	**85.6**
weaponmadeincountry	88.9	87.0	**87.9**	44.4	43.5	44.0	84.7	82.9	83.8
lakeinstate	90.3	93.0	91.6	84.7	86.6	85.6	91.5	93.6	**92.5**
Average all 11 predicates	84.0	83.8	83.9	73.7	73.5	73.6	86.5	86.3	**86.4**

Results and Analysis. We report the performance for each of the three methods in Table 1; the frequency-based, the graph-based, and the combined approach. As expected, we find that the most frequent entity baseline shows strong results. In contrast, the graph-based method shows an overall F_1-measure of only 73.6 compared to 83.9 for the baseline. Our combining approach, however, improves over the most frequent entity baseline by 2.9% w.r.t. average F_1, which is notably a difficult competitor for unsupervised and knowledge-rich methods.

When analyzing the results in detail, we find the combined approach to improve the F_1-measure for all but two NELL predicates. In contrast, the graph-based approach, which does not take into account any information about the term interpretation frequencies, has a great variation in performance: for instance in *actorstarredinmovie*, F_1 increases from 81.3 to 90.5, but for *weaponmadeincountry*, it decreases by 50%, the latter meaning that the graph-based method selects very often highly related, but incorrect subject-object pairs. Analyzing this different performances more detailed, we attribute the improvement to the fact that the underlying knowledge base had sufficient relatedness evidence favoring the likelihood of the correct candidate pairs. For example for *actorstarredinmovie*(morgan freeman, seven), two possible candidate pairs (out of many others) with their probabilities are as follows:

$$(\texttt{dbp:Morgan\_Freeman}, \texttt{dbp:Seven\_Network}) \quad P_{prior} = 0.227; \quad P_{graph} = 0.074$$
$$(\texttt{dbp:Morgan\_Freeman}, \texttt{dbp:Seven\_(film)}) \quad P_{prior} = 0.172; \quad P_{graph} = 0.726$$

With the most frequent entity method, we would have selected the former pair, given its higher prior probability of $P_{prior} = 0.227$. However, the graph-based method captures the relatedness, as DBPEDIA contains the directly connecting edge `dbo:starring` and thus rightly selects the later pair. In other cases, as observed often with *personleadsorganization* and *weaponmadeincountry*, a low prior probability was complemented with a semantic relatedness, thus a high P_{graph},

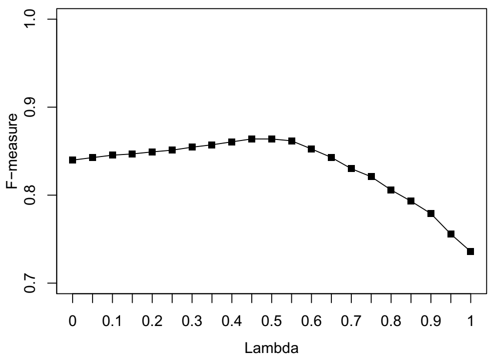

Fig. 1. Effect of Lambda (λ) on the average F_1 score

thereby making a highly related, but incorrect subject-object-combination candidate more likely than the correct one. Consequently, the graph-based approach by itself lowers the performances, relative to the baseline.

The fact that our combined method outperforms both the other approaches indicates that the linear combination of the two probabilities effectively yields in selecting the better of the two methods for each NELL triple. However, in addition to this effect, we observe that our combined approach also finds the correct mapping in cases where both, the frequency-based and the graph-based approach fail individually. Giving one example from the data, for the triple *teamplaysagainstteam(hornets, minnesota timberwolves)*[2], the frequency-based approach disambiguates it to the pair `(dbp:Hornet, dbp:Minnesota_Timberwolves)`, which is incorrect, as `dbp:Hornet` is an insect. But the graph-based approach also disambiguates wrongly to the pair (`dbp:Kalamazoo_College`, `Minnesota_Timberwolves`), even though it discovers a very specific path in DBPEDIA between subject and object in this pair, via the intermediate entity `dbp:David_Kahn_(sports_executive)`. The gold standard pair, (`dbp:New_Orleans_Pelicans, dbp:Minnesota_Timberwolves`), however, gets selected by the combined approach, which combines the medium high prior probability and a medium high relatedness originating from the fact that both instances are connected by `yago:YagoLegalActor`. Not that this last information originates from DBPEDIA and its unsupervised graph weighing method, not from the NELL predicate *teamplaysagainstteam.*

Last, we report on the robustness of our combined approach with respect to the parameter λ, even though giving equal weight to both methods, thus setting λ to 0.5, seems to be a natural choice. Figure 1 shows the F_1-measure for $\lambda \in [0; 1]$. Note that $P_{joint} = P_{graph}$, when $\lambda = 1$ and $P_{joint} = P_{prior}$, when $\lambda = 0$. We observe a clear peak at $\lambda = 0.5$, which confirms our initial choice.

[2] "hornets" refers to `dbp:New_Orleans_Pelicans`, formerly the New Orleans Hornets.

4 Conclusions and Future Work

We addressed the linking of ambiguous NELL subject and object terms to their unique DBPEDIA entities. We studied the effectiveness of a simple baseline which uses the most frequent instance, as well as a knowledge-based approach which exploits DBpedia as a weighted graph. Our contribution is the combination of these two approaches, which outperforms the individual methods at a high level of 86.4 for the F_1-measure. In contrast to other approaches, our method does not require any learning or parameter tuning and the high performance is achieve without using any NELL predicate information. Essentially, we overcome the lack of contextual information in OIE triples by complementing it with existing background knowledge from the target ontology.

As part of future work, we will incorporate the property information from NELL to improve the entity disambiguation. Ultimately, we aim for a NELL predicate disambiguation to DBPEDIA that resolves complex one-to-many property mappings, which could be extracted from those paths currently selected within the cheapest path computation of the graph-based entity linking approach.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics 7(3) (2009)
2. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.: Toward an architecture for never-ending language learning. In: Proc. of AAAI 2010 (2010)
3. Dutta, A., Niepert, M., Meilicke, C., Ponzetto, S.P.: Integrating open and closed information extraction: Challenges and first steps. In: Proc. of the ISWC 2013 NLP and DBpedia Workshop (2013)
4. Etzioni, O.: Search needs a shake-up. Nature 476(7358), 25–26 (2011)
5. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in KnowItAll (Preliminary results). In: Proc. of WWW (2004)
6. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: Proc. of AAAI 2006 (2006)
7. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proc. of EMNLP 2011 (2011)
8. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding light on the web of documents. In: Proc. of I-Semantics (2011)
9. Mihalcea, R., Csomai, A.: Wikify! Linking documents to encyclopedic knowledge. In: Proc. of CIKM 2007 (2007)
10. Schuhmacher, M., Ponzetto, S.P.: Knowledge-based graph document modeling. In: Proc. of WSDM 2014 (2014)
11. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A core of semantic knowledge. In: Proc. of WWW 2007 (2007)
12. Zhang, Z., Gentile, A.L., Ciravegna, F.: Recent advances in methods of lexical semantic relatedness – a survey. Natural Language Engineering 1(1), 1–69 (2012)