AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Data

Ricardo Usbeck[♡♠] and **Axel-Cyrille Ngonga Ngomo**[♡] and **Michael Röder**[♡♠] and **Daniel Gerber**[♡] and **Sandro Athaide Coelho**[♣] and **Sören Auer**[♢] and **Andreas Both**[♠] [1]

Abstract. Over the last decades, several billion Web pages have been made available on the Web. The ongoing transition from the current Web of unstructured data to the Data Web yet requires scalable and accurate approaches for the extraction of structured data in RDF (Resource Description Framework) from these websites. One of the key steps towards extracting RDF from text is the disambiguation of named entities. We address this issue by presenting AGDISTIS, a novel knowledge-base-agnostic approach for named entity disambiguation. Our approach combines the Hypertext-Induced Topic Search (HITS) algorithm with label expansion strategies and string similarity measures. Based on this combination, AGDISTIS can efficiently detect the correct URIs for a given set of named entities within an input text.

1 Introduction

The vision behind the Data Web is to provide a new machine-readable layer to the Web, in which the content of Web pages is annotated with structured data (e.g., RDFa [1]). Most of these websites are unstructured in nature. Realizing the vision of an easy to use and up-to-date Data Web thus requires scalable and accurate natural-language-processing approaches that allow extracting RDF from such unstructured data. One central task during this process is named entity disambiguation (NED, also called entity linking). Current Named Entity Disambiguation (NED) approaches suffer from two major drawbacks: they perform poorly on Web-documents and rely on exhaustive data mining methods or algorithms with non-polynomial time complexity.

In this paper, we address these drawbacks by presenting AGDISTIS, a novel NED approach and framework. AGDISTIS computes results by combining the HITS algorithm [4] with label expansion and string similarity measures. A demo of our approach (integrated into the Named Entity Recognition framework FOX) can be found at `http://fox.aksw.org`.

2 The AGDISTIS Approach

Our approach to NED consists of three main phases: Given an input text T and a named entity recognition function (e.g., [6]), we begin by retrieving all named entities from the input text. Thereafter, we aim to detect candidates for each of the detected named entities. To this end, we apply several heuristics and make use of known surface forms [5] for resources from the underlying KB. The set of candidates generated by the first step is used to generate a disambiguation graph. Second, we rely on a graph search algorithm which retrieves context information from the underlying KB. Finally, we employ the HITS algorithm to the context graph to find authoritative candidates for the discovered named entities. We assume that the resources with the highest authority values represent the correct candidates. All algorithms in AGDISTIS have a polynomial time complexity, leading to AGDISTIS also being polynomial in time complexity. In the following, we present each of the steps of AGDISTIS in more detail.

Candidate Detection. In order to find the correct disambiguation for a certain set of named entities, we first need to detect candidate resources in the KB. We begin by creating an index comprising all labels of each resource. Our approach can be configured to use any set of properties as labeling properties. For our experiments, we only considered `rdfs:label` as labeling property. In addition, our approach can make use of known *surface forms* for each of the resources in case such knowledge is available [5]. These are strings that are used on the Web to refer to given resources. Surface forms are added to the set of available labels for each resource. In this paper, we do not consider abbreviations although these could be easily included in the process by adding further labels into the KB (e.g., via WordNet[2]).

Next to searching the index we apply a *string normalization* approach and an *expansion policy* to the input text: The string normalization is based on eliminating plural and genitive forms, removing common affixes such as postfixes for enterprise labels and ignoring candidates with time information (years, dates, etc.) within their label. For example, the genitive `New York's` is transformed into `New York`, the postfix of `Microsoft Ltd.` is reduced to `Microsoft` and the time information of `London 2013` is ignored. Our *expansion policy* is a time-efficient approach to coreference resolution [7], which plays a central role when dealing with text from the Web. In web and news documents, named entities are commonly mentioned in their full length the first time they appear, while the subsequent mentions only consist of a substring of the original mention due to the brevity of most news data. For example, a text mentioning Barack Obama's arrival in Washington D.C.

[1] ♡ University of Leipzig, Germany , ♠ R & D, Unister GmbH, Leipzig, Germany, ♣ Federal University of Juiz de Fora, Brazil, ♢ University of Bonn & Fraunhofer IAIS, Germany
email: {usbeck|ngonga}@informatik.uni-leipzig.de

[2] `http://wordnet.princeton.edu/`

will commonly contain `Barack Obama` in the first mention of the entity and use strings such as `Obama` or `Barack` later in the same text. We implement this insight by mapping each named entity label (e.g., `Obama`) which is a substring of another named entity label that was recognized previously (e.g., `Barack Obama`) to the same resource (i.e., `dbr:Barack_Obama`). If there are several possible expansions, we choose the shortest as a fast coreference resolution heuristic for web documents. Without the expansion policy AGDISTIS suffers from a loss of accuracy of $\approx 4\%$.

Additionally, AGDISTIS can be configured to fit named entities to certain domains to narrow the search space, e.g., for targeting DBpedia and persons, named entity types would be `dbo:Person, foaf:Person`. Obviously, these classes can be altered by the user as required to fit his purposes.

In its final step, our system compares the heuristically obtained label with the label extracted from the KB by using *trigram similarity* which is a n-gram similarity with $n = 3$.

Computation of Optimal Assignment. Given a set of candidate nodes, we begin the computation of the optimal assignment by constructing a disambiguation graph G_d with search depth d. To this end, we regard the input knowledge base as a directed graph $G_K = (V, E)$ where the vertices V are resources of K, the edges E are properties of K and $x, y \in V, (x, y) \in E \Leftrightarrow \exists p : (x, p, y)$ is an RDF triple in K. Given the set of candidates C, we begin by building an initial graph $G_0 = (V_0, E_0)$ where V_0 is the set of all resources in C and $E_0 = \emptyset$. Starting with G_0 we extend the graph in a breadth-first search manner. Therefore, we define the extension of a graph $G_i = (V_i, E_i)$ to a graph $\rho(G_i) = G_{i+1} = (V_{i+1}, E_{i+1})$ where $i = 0, \ldots, d$ as follows:

$$V_{i+1} = V_i \cup \{y : \exists x \in V_i \wedge (x, y) \in E\} \quad (1)$$

$$E_{i+1} = \{(x, y) \in E : x, y \in V_{i+1}\} \quad (2)$$

We iterate the ρ operator d times on the input graph G_0 to compute the initial disambiguation graph G_d.

After constructing the disambiguation graph G_d, we need to identify the correct candidate node for a given named entity. Using the graph-based HITS algorithm we calculate authoritative values x_a, y_a and hub values x_h, y_h for all $x, y \in V_d$. We initialize the authoritative and hub values as follows:

$$\forall x \in V_d, x_a = x_h = \frac{1}{|V_d|}. \quad (3)$$

Afterwards, we iterate k times the following equations:

$$x_a \longleftarrow \sum_{(y,x) \in E_d} y_h, \quad y_h \longleftarrow \sum_{(y,x) \in E_d} x_a. \quad (4)$$

We choose k according to [4], i.e., 20 iterations, which suffice to achieve convergence in general. Afterwards, we identify the most authoritative candidate C_{ij} among the set of candidates C_i as correct disambiguation for a given named entity N_i. When using DBpedia as KB and C_{ij} is a redirect, AGDISTIS uses the target resource. As can be seen, we calculate the optimal assignment solely by using polynomial time complex algorithms.

3 Preliminary Results

We carried out a preliminary evaluation of our approach against state-of-the-art approaches (i.e., TagMe 2 [2], AIDA [3] and DBpedia Spotlight [5]) on the MSNBC dataset. There, we achieve 76.1% F-measure and outperform the state of the art by up to 29.5% F-measure. Preliminary results provided the following early findings: (1) Varying the search depth d does not significantly improve F-measure because within the underlying documents there are many similar named entities forming a shallow semantic background. (2) However, using only string similarity measures ($d = 0$) results in lower F-measure. (3) The expansion policy can have considerable knock-on effects: Either the first entity and its expansions are disambiguated correctly or the wrong disambiguation of the first entity leads to an avalanche of false results. (4) Using $n = 1, 2, 4$ as n-gram similarity has been proven to perform worse than using trigram similarity, i.e., $n = 3$.

4 Conclusion

We presented AGDISTIS, a novel, LOD-background-knowledge-based named entity disambiguation approach. By combining the scalable HITS algorithm and breadth-first search with linguistic heuristics, we were able to precisely disambiguate a wide range of named entities. We see this work as the first step in a larger research agenda. Based on AGDISTIS, we aim to develop a new paradigm for realizing NLP services, which employ community-generated, multilingual and evolving LOD background knowledge.

Acknowledgments

This work has been supported by the ESF and the Free State of Saxony and the FP7 project GeoKnow (GA No. 318159).

References

[1] B. Adida, I. Herman, M. Sporny, and M. Birbeck, 'Rdfa 1.1 Primer', Technical report, World Wide Web Consortium, http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/, (June 2012).

[2] Paolo Ferragina and Ugo Scaiella, 'Fast and accurate annotation of short texts with wikipedia pages.', *IEEE software*, **29**(1), (2012).

[3] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum, 'Robust Disambiguation of Named Entities in Text', in *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*, pp. 782–792, (2011).

[4] Jon M. Kleinberg, 'Authoritative sources in a hyperlinked environment', *J. ACM*, **46**(5), 604–632, (September 1999).

[5] Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer, 'Dbpedia spotlight: Shedding light on the web of documents', in *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, (2011).

[6] Axel-Cyrille Ngonga Ngomo, Norman Heino, Klaus Lyko, René Speck, and Martin Kaltenböck, 'Scms–semantifying content management systems', in *The Semantic Web–ISWC 2011*, 189–204, Springer, (2011).

[7] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum, 'Large-scale cross-document coreference using distributed inference and hierarchical models', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 793–803. Association for Computational Linguistics, (2011).