DAGOBAH: Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data

Viet-Phi Huynh[1], Jixiong Liu[1], Yoan Chabot[1], Frédéric Deuzé[1], Thomas Labbé[1], Pierre Monnin[1], and Raphaël Troncy[2]

[1] Orange, France
yoan.chabot@orange.com
[2] EURECOM, Sophia Antipolis, France
raphael.troncy@eurecom.fr

Abstract. The annotation of tabular data is a strategic issue for companies as it provides an automatic understanding of a data structure that is at the heart of many services and products. In recent years, Orange, in collaboration with EURECOM, has developed tools for pre-processing and semantic annotation of tables to meet industrial challenges. In this paper, we present the latest evolutions of the DAGOBAH system developed in the framework of the SemTab2021 challenge. In particular, optimisations in the lookup mechanisms and new techniques for studying the context of the target knowledge graph nodes have enabled us to obtain very promising results. To accelerate the adoption of STI solutions within the enterprise, this paper also presents the deployment of these algorithms via the TableAnnotation API and DAGOBAH UI.

Keywords: Semantic Table Interpretation · DAGOBAH · SemTab

1 Introduction

Over the last three years and thanks to the participation in the SemTab challenge, the annotation tools developed by Orange in collaboration with EURECOM have reached a degree of maturity high enough to start addressing industrial use cases in targeted areas. As a multinational company working in very different fields (telecommunications but also cybersecurity, multimedia content, etc.), Orange produces very large volumes of heterogeneous tabular data in a daily basis. These tables contain an important part of the company's knowledge and are used extensively in many services and products. Therefore, automatically understand them by matching their elements with entities of Knowledge Graphs (either encyclopedia KGs like Wikidata/DBPedia or specific domain/enterprise ones) is strategic for a company. Indeed, STI tools are of interest for many use cases and business areas including 1) Data governance where STI tools can generate semantic annotations on datasets, hence, boosting the capabilities of indexing tools and data catalogs with semantics; 2) Data science projects where users can leverage preprocessing and annotation tools to better clean, understand, reconcile and enrich datasets; 3) Knowledge capitalization where STI can help to make the dormant knowledge actionable i.e. structure it and make it usable through Q/A engines for example.

To meet these industrial challenges and in the context of the SemTab2021 challenge, we have improved the entity scoring algorithm, the core of DAGOBAH SL 2020 system [3]. This paper presents the following contributions:

- An enhancement of the indexing and entity matching strategies to improve the lookup quality as well as lookup coverage.
- A better representation and disambiguation of the entities by exploiting more efficiently their contexts in the KG.
- An improved and flexible entity scoring considering multiple factors which leverage both local information and global table information.

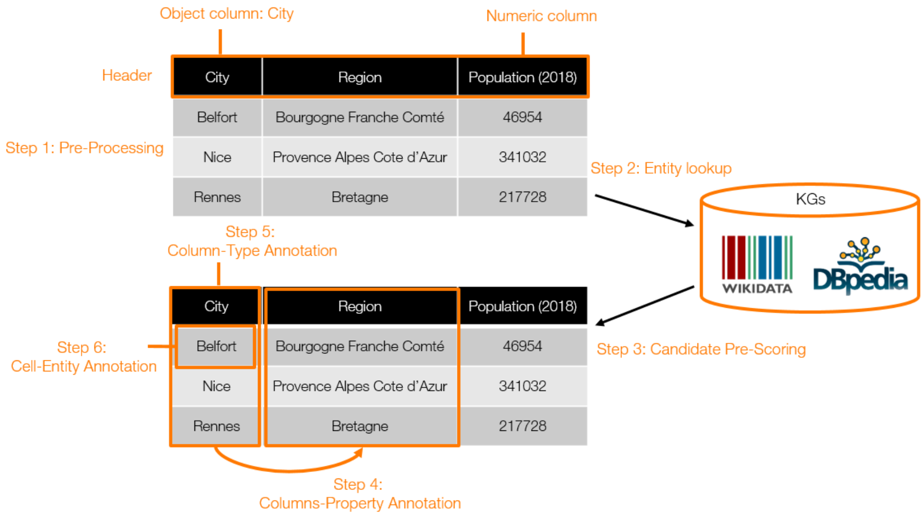
The remainder of this paper is organized as follows. In Section 2, DAGOBAH SL 2021 and the improvements made to meet the specificities of the SemTab2021 challenge are presented. The results of the experimental evaluation, the insights gained from it as well as the challenge leaderboard are then reported in Section 3. Section 4 presents the efforts made around the usability of DAGOBAH within Orange, in particular via the TableAnnotation API and DAGOBAH UI. Finally, Section 5 discusses the perspectives on the SemTab challenge and the adoption of STI solutions in the enterprise.

2 DAGOBAH SL 2021: Optimised Lookup, Exploration of Knowledge Graphs and Flexible Entity Scoring

DAGOBAH provides an end-to-end annotation for relational tables leveraging a KG, i.e, Wikidata. The system's processing pipeline consists of four sequential steps depicted in Figure 1. Given a relational table as input, the pre-processing (Section 2.1) first extract table metadata and determine the annotation targets. The entity lookup module then collects candidates from the KG for each target table cell (Section 2.2). The pre-scoring module flexibly evaluates each candidate with a confidence score considering multiple factors described in Section 2.3. Next, the Columns-Property Annotation (CPA) and Column Type Annotation (CTA) are carried out to annotate single columns and column pairs with types and relations (Section 2.4). The entity scores are then calibrated taking the CTA and CPA into account to carry out the Cell-Entity Annotation (CEA)(Section 2.4).

2.1 Table Pre-Processing

In real use cases, it is complex to annotate a table with little or no prior knowledge about its structure and content. Therefore, having the table preprocessed



City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

Fig. 1: Overview of the DAGOBAH annotation workflow.

can facilitate the later annotation. DAGOBAH generates metadata about a table via four main tasks[3]: orientation detection, header detection, key column detection (only single key column currently supported) and column primitive typing. The preprocessing step was particularly helpful for the annotation of the BioDivTables and GitTables corpora (round 3). The proposed solution leverages the typing of a column, i.e. named entities (Location, Organization, Country, etc.), unit entities (Distance, Speed, Temperature, etc.) or miscellaneous entities (Email, URL, IP Address,etc.) to find a mapping with Schema.org/DBPedia Ontology (GitTables) or to discriminate object/literal columns (BioDivTables). For the later, a column likely contains literal value if its typing belongs to numerical/date-time entity, unit entities, or misc entities. Otherwise, it is considered as an object column containing entities mentions on which the entity lookup needs to be triggered.

2.2 Entities Lookup

The pre-processing helps to identify table's columns that are potentially lookupable (in the context of SemTab, the target cells are given and, therefore, this feature is not used). Given a cell e_m in such column, the entity lookup step queries a target KG and collects a set of relevant candidate entities $\mathcal{E}_c$. The lookup service of DAGOBAH, based on Elasticsearch, currently supports 2 KGs: Wikidata and DBPedia.

- For Wikidata entities, the service collects items and properties (respectively identified by QID and PID) together with their labels and aliases in all

[3] This toolkit will be the subject of a future paper.

languages. To increase the coverage of lookups operations, the aliases of each entity are enriched with the values of 11 additional properties including P2561 (name), P1705 (native label) or P742 (pseudonym).
- For DBPedia entities, the service collects english resources (prefixed by dbpedia.org/resource) with their labels in all languages. As before, 25 additionnal alias properties values are collected to enhance the coverage of the system including ”abbreviation”, ”birthName” or ”originalTitle”. In addition, the labels and aliases of all redirected entities are also included.

We average the character-based and token-based edit distances[4] to evaluate the similarity between a cell mention and the set of labels of the candidate entities. This helps to solve the out of order problem where a string have many different order of its substrings (i.e., ”Elon Musk” and ”Musk Elon”).

2.3 Candidate Pre-Scoring

This section presents the entity scoring algorithm which is the core of DAGOBAH SL. As part of it, the pre-scoring step evaluates the relevance of a candidate entity $e_c \in \mathcal{E}_c$ of table cell e_m with a preliminary score. This algorithm is based on the DAGOBAH SL 2020's scoring ([3], Section 3.2) on which is employed the same definition of score function $Sc(e_c, e_m)$ as well as terminologies.

Equation (1) re-states the context score $Sc(e_c, e_m)$ which was the main focus of the 2021 system improvements:

$$Sc_{context}(\mathcal{N}_{table}(e_m), \mathcal{N}_{graph}(e_c)) = \overline{\mathcal{N}_s} = \frac{\sum_i w_i \times sn_i}{\sum_i w_i} \tag{1}$$

where $\mathcal{N}_{table}(e_m)$ is the set of neighboring cells in the same row as e_m. $\mathcal{N}_{graph}(e_c)$ is the set of neighboring elements of e_c in the KG[5]. $\mathcal{N}_s$ is a set which contains neighborhood matching score sn_i for each neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ w.r.t $\mathcal{N}_{graph}(e_c)$. DAGOBAH SL 2021 solves two issues related to the calculation of context score of DAGOBAH SL 2020:

- First, each context score component sn_i in $\mathcal{N}_s$ is evaluated expensively by iterating over all the context nodes of e_c in $\mathcal{N}_{graph}(e_c)$ to find the best matching node. For example, in the table in Figure 1, considering the annotation of cell “Belfort” which has a Wikidata candidate entity Q171545. To check whether the neighboring cell ”Bourgogne Franche Comté” is a context of Q171545, we have to browse $\sim$ 1000 nodes of $\mathcal{N}_{graph}$(Q171545) (Figure 2.a) including ”France” (Q142), ”Paul Faivre” (Q3371185), etc. and perform the comparison on each node. The performance bottleneck arises when it comes to scoring a generic entity of millions edges in KG, i.e. ”France” (Q142).
- Second, the neighboring elements of e_c are nodes only one step away from e_c in the KG. Consequently, a neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ matching with

[4] https://github.com/seatgeek/thefuzz

[5] Neighboring elements are nodes connected to e_c via predicate paths on KG.

a node located two hops away from e_c in KG is not considered as context of e_c. For example, given the one-depth graph centered around Q171545 (”Belfort”) (Figure 2.a), it is wrong saying that ”Bourgogne Franche Comté” has no relation with Belfort city since “Bourgogne Franche Comté” is the region of “Territoire de Belfort” departement whose capital is “Belfort”.

Given those issues, DAGOBAH SL 2021 improves both the efficiency and the expressivity of the context score by withdrawing the exhaustive scoring and exploiting more expressive contexts for the graph of an entity via 2-hop predicate paths. To illustrate these improvements, we re-use the table in Figure 1 and the current annotation target described in the two issues above. That is, a score (sn_i) is given to Q171545 (e_c), a candidate entity of ”Belfort” (e_m) which has a neighboring cell ”Bourgogne Franche Comté” (n_i).

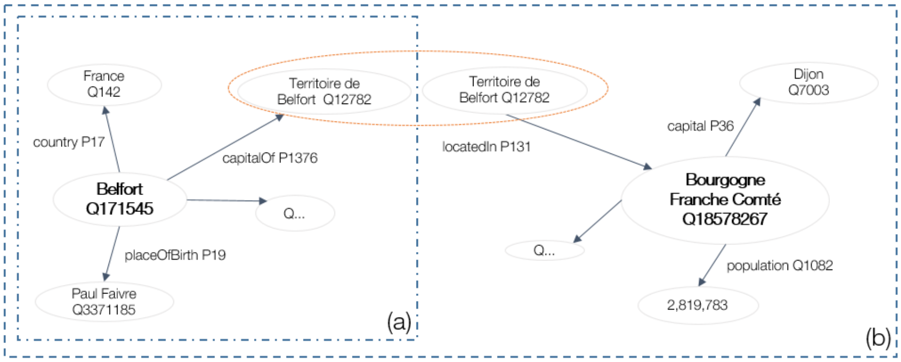


Fig. 2: KG graph context of entity Belfort (Q171545) in Wikidata: (a) One-hop graph context centered around Q171545. (b) Graph context is expanded by subgraphs intersection.

Exploiting the Context of Knowledge Graph Entities The context score component ns_i in Equation (1) indicates whether the corresponding neighboring cell n_i matches with a node on the graph $\mathcal{N}_{graph}(e_c)$ of target entity e_c. Loosely speaking, computing ns_i resorts to searching, on the graph $\mathcal{N}_{graph}(e_c)$, a candidate entity for the neighboring cell n_i and assess its closeness with n_i. For example, on two-depth graph $\mathcal{N}_{graph}(e_c$=Q171545) (Figure 2.b), Q18578267 ($e_{c'}$) is a candidate of n_i = ”Bourgogne Franche Comté”. Given this observation, we propose an effective way to determine ns_i as follows. The entity lookup step in Section 2.2 gives not only the candidate entities $\mathcal{E}_c$ for target cell e_m but also the candidate entities $\mathcal{E}_{c'}$ for the neighboring cell n_i. We now verify if a candidate entity $e_{c'} \in \mathcal{E}_{c'}$ of n_i is part of neighbor elements of target candidate entity e_c; in other words, $e_{c'}$ is connected to e_c by a predicate path in

KG, i.e, Q171545 (e_c) traverses along "capitalOf (P1376), Territoire de Belfort (Q12782), locatedIn (P131)" to Q18578267 ($e_{c'}$) (Figure 2.b). In that case, the score ns_i is simply calculated by comparing the neighboring cell label n_i and the matching node $e_{c'}$ avoiding trivial comparisons with other nodes on $\mathcal{N}_{graph}(e_c)$. A method of sub-graph extraction is used to find predicate paths over a node pair (e_c, $e_{c'}$) in a graph, at reasonable expense. Particularly, for each node in the pair, a sub-graph $\mathcal{G}$ centered around that node is extracted. Merging the two sub-graphs $\mathcal{G}_{e_c}$ of e_c and $\mathcal{G}_{e_{c'}}$ of $e_{c'}$ allows to find predicate paths, if exists, linking e_c to $e_{c'}$. That is, if an intermediate node i is present in both $\mathcal{G}_{e_c}$ and $\mathcal{G}_{e_{c'}}$, the paths pointing to i in two sub-graphs are concatenated, such as "capitalOf (P1376), locatedIn (P131)". Only 1-depth sub-graphs, making two-hop path the longest path that can be retrieved for (e_c, $e_{c'}$), i.e., $e_c \xrightarrow{p_1}$ i $\xrightarrow{p_2} e_{c'}$. Intuitively, sub-graphs intersection allows to enrich the information about an entity by including not only direct neighboring nodes but also indirect nodes two step away from it. This richer representation of e_c increases the chance for a neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ to match, making the context score more precise. We argue that for a node pair, the connecting paths longer than 2-hop path, which are very often noisy and meaningless, can have a negative impact on the context score.

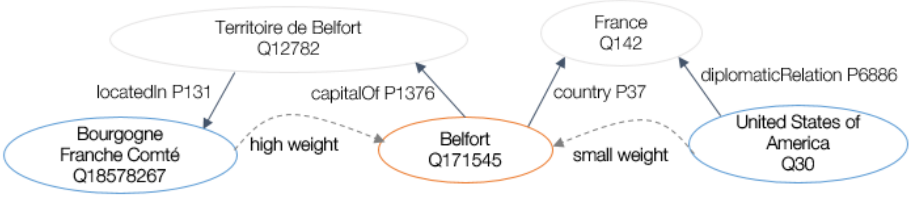


Fig. 3: Neighboring nodes of Belfort (Q171545) contribute differently to its information content.

Soft Context Scoring The context score components are weighted to obtain the ultimate score of an entity (Equation 1), implying that each neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ contributes differently to the annotation of the target cell e_m (e.g. cells containing entities should be more important than literal ones).

$$w_i = se_i \tag{2}$$

where $se_i = 1.0$ if the corresponding neighboring cell n_i refers to an entity, and 0.15 if n_i is a literal value (date, measurement with/without unit, number, etc.). As there is a lack of literal value disambiguation methods (i.e. date time normalization, unit detection/normalization/conversion), it can be risky to give too much importance to literal contexts in the entity score. DAGOBAH SL 2021 weights the importance of the different types of neighbouring cells via additional mechanisms:

- A neighboring cell that is closer to the extreme left side of the table has more chance to be a meaningful context of target cell.

$$w_i = \frac{se_i}{\sqrt{d(col_i)+1}} \tag{3}$$

  where $d(col_i)$ is the distance between column col_i associated with w_i and the first object column.
- A neighboring column col_i that is highly connected to the target column should have its cells n_i to be taken more into account than less connected neighboring column. The connectivity of a neighboring column (cnt_{col_i}) w.r.t target column is expressed by the highest occurrence of a relation possibly found behind the two. The context weight w_i is then updated with:

$$w_i = \frac{se_i}{\sqrt{d(col_i)+1}} \times cnt_{col_i} \tag{4}$$

- On the graph $\mathcal{N}_{graph}(e_c)$ of target entity e_c, neighboring nodes provide different information content. One node can be semantically closer to e_c than the other. For e.g., considering the 2-depth graph of Q171545 Belfort (Figure 3), it is clear that the node "Bourgogne Franche Comté (Q18578267)" is more relevant than the node "United States of America (Q30)" since the path *Belfort* $\xrightarrow{capitalOf}$ *Territoire de Belfort* $\xrightarrow{localtedIn}$ *Bourgogne Franche Comté* is much more informative than the path *Belfort* $\xrightarrow{country}$ *France* $\xleftarrow{diplomaticRelation}$ *United States of America*. To quantify the so-called truth value $\tau(e_{c'})$ [2] of a neighboring node $e_{c'}$ or the discriminative capacity of the associated path $\tau(\{p_1, p_2\})$, i.e. $e_c \xrightarrow{p_1} i \xrightarrow{p_2} e_{c'}$, we rely on the *generality* $g(i)$ of the intermediate node i:

$$\tau(e_{c'}) = \tau(\{p_1, p_2\}) = \frac{1}{1 + log(g(i))} \tag{5}$$

  where the *generality* $g(i)$ is the number of incoming and outcoming edges for node i in the KG. Note that the direct neighboring node (or 1-hop predicate path) always get highest truth value 1.0. The context weight w_i (Equation 4) is then updated by the discriminative capacity of the most frequent relation r between target column and neighboring column:

$$w_i = \frac{se_i}{\sqrt{d(col_i)+1}} \times cnt_{col_i} \times \tau(r). \tag{6}$$

2.4 Annotation Tasks

The CPA task involves finding the most suitable semantic relation r between a pair of ordered columns $\{\text{head}, \text{tail}\}$. A majority voting strategy is adopted which relies on the number of occurrences and accumulated confidence score over rows of the relation r. The reader can refer to [3] for additional details on how to calculate the scores. Note that, according to Entity Scoring section, a relation between an entity pair can be one-hop (i.e. $s \xrightarrow{p} o$) , unidirectional 2-hop (i.e. $s \xrightarrow{p1}\xrightarrow{p2} o$), bidirectional 2-hop (i.e. $s \xrightarrow{p1}\xleftarrow{p2} o$), CPA annotations outputed by our system can thus be one of those.

The CTA tasks aims to identify the most representative type for a target column. From the candidate entity set of row cells in this column, the types of each entity are collected and a majority voting strategy is used to determine the most precise type. [3] provides more details on the type enrichment step as well as score calculation.

The CEA task refers to selecting, for a table cell e_m, the most relevant entity e_c among a set of candidate entity $\mathcal{E}_c$ of e_m retrieved from KG. This step is based on the entity pre-scoring but also leverage the information given by CTA and CPA to compute the final score of entity e_c. The preliminary score of candidate entity e_c calculated from Entity Pre-scoring step only considers its local information, that is the row it belongs to. To exploit table's global information spreading over rows and columns, we leverage the column type (CTA) and column pair relation (CPA), which inherently aggregate inter-row, inter-column information. In more details, the score $Sc(e_c, e_m)$ of cell entity e_c is calibrated as follows:

$$Sc(e_c, e_m) = \frac{(Sc(e_c, e_m) + \alpha \times score_{CTA} + \beta \times \overline{score_{CPA}})}{1 + \alpha + \beta} \tag{7}$$

where $score_{CTA}$ is the score of the column type associated with e_c, $\overline{score_{CTA}}$ is the aggregated score of column pair relations attached to the column of e_c. To encourage (resp. discourage) a frequent (resp. unusual) CTA/CPA to participate in the update of $Sc(e_c, e_m)$, a coefficient α (resp. β) is employed and defined as $\frac{occurence(CTA)}{2}$ (resp. $\frac{occurence(CPA)}{2}$). Note that the *occurrence* of CTA/CPA is divided by 2 to always prioritize the preliminary $Sc(e_c, e_m)$ during the update.

3 Experiments

3.1 Settings

To evaluate the behavior of different types of entity's graph context (one-hop, two-hop) as well as the soft context scoring (Section 2.3), four system settings are considered for the experiments:

- Setting 1. The context score of an entity is calculated using only its one-hop neighboring graph. The contribution of a context follows Equation 2 in which the weights are fixed to 1.0 for entity contexts and 0.15 for literal contexts.
- Setting 2. The context score of an entity is calculated using its two-hop neighboring graph. The configuration of the importance of a context also follows Equation 2, but with a weight of 0.25 for 2-hop entity context, instead of 1.0.

Table 1: Evaluation results of annotation settings and Overall results of DAGOBAH system on rounds 1, 2, 3 of Semtab2021 challenge. “F1” stands for F1-score, “P” stands for Precision. Complete results will appear in final paper version.

			CTA		CEA		CPA	
			F1	P	F1	P	F1	P
Evaluation	Round 1-WDTables (132/180 lightest tables)	Setting 1	x	x	0.175	0.812	x	x
		Setting 2	x	x	0.179	0.831	x	x
		Setting 3	x	x	0.185	0.860	x	x
		Setting 4	x	x	0.185	0.860	x	x
	Round 2-HardTables	Setting 1	x	x	0.975	0.976	x	x
		Setting 2	x	x	0.976	0.976	x	x
		Setting 3	x	x	0.976	0.976	x	x
		Setting 4	x	x	0.976	0.976	x	x
LeaderBoard	Round 1-WDTables	Setting 2*	0.832	0.832	0.923	0.923	x	x
	Round 1-DBPTables	Setting 2*	0.422	0.424	0.945	0.946	x	x
	Round 2-BioTables	Setting 3	0.916	0.916	0.970	0.970	0.899	0.899
	Round 2-HardTables	Setting 3	0.976	0.976	0.975	0.976	0.996	0.996
	Round 3	TBD	TBD	TBD	TBD	TBD	TBD	TBD

- Setting 3. The context score of an entity is calculated using its two-hop neighboring graph. The soft context scoring (Equation 6) is used to estimate the information gain of a context using multiple factors. This setting aims to check if richer contexts and stricter scoring means better annotation.
- Setting 4. This context is similar to Setting 3 but only unidirectional 2-hop predicate paths ($e_1 \xrightarrow{p_1}$ i $\xrightarrow{p_2} e_2$ and $e_1 \xleftarrow{p_1}$ i $\xleftarrow{p_2} e_2$) in entity’s graph are employed. This setting studies the impact of bi-directional contexts (i.e. $e_1 \xrightarrow{p_1}$ i $\xleftarrow{p_2} e_2$ which are very often less informative or noisy but may be helpful for the disambiguation in some cases.

3.2 Results

The annotation system have been continuously improved during the Semtab2021 challenge. An initial comparison of different annotation settings is shown in Table. 1, section Evaluation. Due to the time constraint, only the CEA results are reported (as it allows to evaluate the system’s improvements made in the new version of our system) on 132/180 lightest tables from round 1’s WDTables corpus and 1750/1750 tables from round 2’s HardTables (more complete results will be given in the final paper version). In our opinion, WDTables and HardTables are, respectively, the most difficult and easiest corpus during rounds 1 and 2.

Our system achieves a very good performance on HardTables and it can be see that using a richer entity graph or a more flexible scoring has no clear gain in this case. This can be explained by the synthetic nature of HardTables corpus in which a table is almost fully represented in the target KG. Meanwhile, in WDTables, different behaviors are observed across systems[6]. The settings 2, 3 and 4 are more precise than the setting 1, implying that meaningful graph contexts of an entity can be exploited more-than-one step away from it. The effectiveness of soft context scoring is then demonstrated through the out-performance of setting 3 and 4 over 2. Finally, the observation that setting 3 behaves similarly to 4 may lead to two conclusions. First, uni-directional contexts ($e_1 \xrightarrow{p_1} \mathrm{i} \xrightarrow{p_2} e_2$ and $e_1 \xleftarrow{p_1} \mathrm{i} \xleftarrow{p_2} e_2$) are more informative than others kinds of context. Using only uni-directional contexts can already obtain equal results. Second, noisy bi-directional contexts (i.e. $\textit{Belfort} \xrightarrow{country} \textit{France} \xleftarrow{diplomaticRelation}$ *United States of America* of form $e_1 \xrightarrow{p_1} \mathrm{i} \xleftarrow{p_2} e_2$) are well controlled in soft context scoring in order not to degrade the overall annotation quality, paving the way for others useful bi-directional contexts to contribute positively to the entity score.

To position our system on the leaderboard of the Semtab2021 challenge, Table. 1, section LeaderBoard provides the annotation scores for CEA, CTA, CPA tasks for rounds 1 and 2[7]. During the round 1, DAGOBAH SL 2021 was not fully updated with all enhancements described in the paper. Indeed, the setting 2* used during round 1 differs from the original Setting 2: it considers only uni-directional predicate paths to avoid the noisy bi-directional ones. In general, the round 2 is easier to annotate because the corpus is generated in a synthetic way, while round 1 exposes high quality manually-curated tables with complex patterns. It can be observed than the CTA, CPA does not work as expected in most corpus despite of the good quality of CEAs. The development of a better strategy for type selection will be the object of future works. Note that, in GitTables, as the table contents are significantly ambigous, even for human, the system described in the paper was not used. Instead, the pre-processing (Section 2.1) was used to extract primitive typings for target columns and our solution then performs a mapping to the target ontologies (DBPedia/Schema.org).

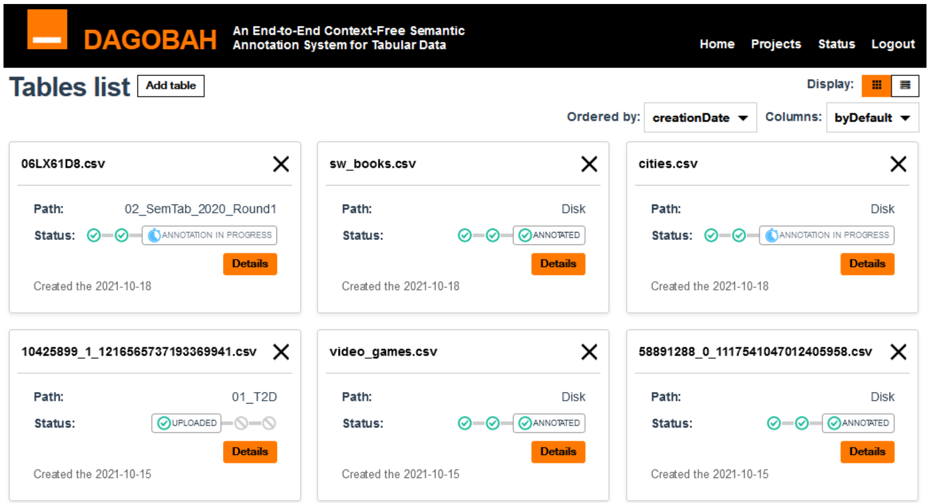
4 Interpreting Tabular Data at Orange

To quickly improve the relevance of DAGOBAH tools, our research team adopted a test & learn approach, by making the annotation algorithms available to other collaborators inside the company very early in the project. Therefore, the versions of DAGOBAH used during the SemTab challenge are made available via an API named TableAnnotation. This RESTful API, deployed on the Orange Developer portal[8], provides access to solutions to pre-preprocess and annotate

[6] Given that all considered systems employ a unique set of entity lookup candidate, we can ignore the low F1 score which is explained by only a portion of corpus (130 lightest tables/180 tables) used during the evaluation and focus on the precision. The final version of the paper will include correct F1 scores.

[7] The results for round 3 will be updated in the final paper version.

[8] https://developer.orange.com



06LX61D8.csv
Path: 02_SemTab_2020_Round1
Status: ANNOTATION IN PROGRESS
Created the 2021-10-18

sw_books.csv
Path: Disk
Status: ANNOTATED
Created the 2021-10-18

cities.csv
Path: Disk
Status: ANNOTATION IN PROGRESS
Created the 2021-10-18

10425899_1_1216565737193369941.csv
Path: 01_T2D
Status: UPLOADED
Created the 2021-10-15

video_games.csv
Path: Disk
Status: ANNOTATED
Created the 2021-10-15

58891288_0_1117541047012405958.csv
Path: Disk
Status: ANNOTATED
Created the 2021-10-15

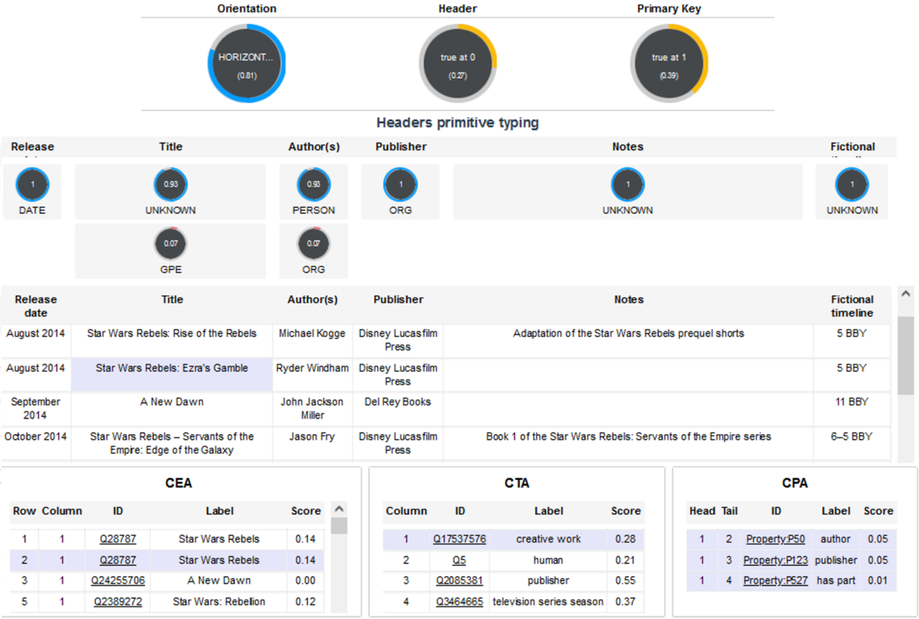
Fig. 4: DAGOBAH UI allows to load in a project a table from the local file system or a preloaded table from a gold standard (e.g. T2D, SemTab, etc.)

tables, as well as to lookup services to disambiguate mentions and get Wikidata/DBpedia entities in return. This API is accessible to all the company's R&D teams as well as to business units that request it (there is plan to extend access to the API to external users in the future). The availability of our system allows the collaborators to understand the interest of STI solutions while allowing the DAGOBAH project team to identify the difficulties associated with their needs, which is a very interesting input for the project roadmap.

The STI algorithms presented in this paper are also promoted in DAGOBAH UI. This interface allows non-developers and non-AI experts to use the TableAnnotation API resources on their tables and visualise the results in an intelligible and ergonomic form. DAGOBAH UI is also a very powerful tool to demonstrate the value of STI within the company or with external prospects. As shown in Figure 4, the interface allows the user to load new tables into their annotation project. The user can then call the preprocessing and semantic annotation tools and visualise the results as seen in Figure 5.

As DAGOBAH UI is a powerful lever for adoption and usability, it will be the subject of many future developments and efforts. Through the STI capabilities presented, DAGOBAH UI is able to map tables with related entities of the KG. This connection between the table and the KG is the basis of several future features:

- The capability to enrich KGs with table elements not present in the KG. The CTA and CPA annotations allow to instantiate in the graph new individuals from unmatched table rows.



Release date	Title	Author(s)	Publisher	Notes	Fictional timeline
August 2014	Star Wars Rebels: Rise of the Rebels	Michael Kogge	Disney Lucasfilm Press	Adaptation of the Star Wars Rebels prequel shorts	5 BBY
August 2014	Star Wars Rebels: Ezra's Gamble	Ryder Windham	Disney Lucasfilm Press		5 BBY
September 2014	A New Dawn	John Jackson Miller	Del Rey Books		11 BBY
October 2014	Star Wars Rebels – Servants of the Empire: Edge of the Galaxy	Jason Fry	Disney Lucasfilm Press	Book 1 of the Star Wars Rebels: Servants of the Empire series	6–5 BBY

CEA

Row	Column	ID	Label	Score
1	1	Q28787	Star Wars Rebels	0.14
2	1	Q28787	Star Wars Rebels	0.14
3	1	Q24255706	A New Dawn	0.00
5	1	Q2389272	Star Wars: Rebellion	0.12

CTA

Column	ID	Label	Score
1	Q17537576	creative work	0.28
2	Q5	human	0.21
3	Q2085381	publisher	0.55
4	Q3464665	television series season	0.37

CPA

Head	Tail	ID	Label	Score
1	2	Property:P50	author	0.05
1	3	Property:P123	publisher	0.05
1	4	Property:P527	has part	0.01

Fig. 5: At the top, information generated by the preprocessing (orientation, header, etc.) as well as the cleaned table are displayed. The bottom is an interactive view with the data table and the CEA, CTA and CPA annotations.

- The ability to enrich the table from the KG. This enrichment can be done at two levels: 1) filling in missing values by using the CEA and CPA annotations and 2) adding new columns by taking advantage of the CEA annotations.
- Interactive visualisations of the target KG, the entities identified by the CEA/CTA/CPA annotation steps and the new triples that can be generated from the table.

5 Discussions

The dataset provided for the Semtab challenge have evolved over the last three years, as the organizers used several tricks to make the matching harder for competitors. New target domains (biomedicine, git data) as well as KGs combination or connate ontologies introduction (Schema.org) also leaded to extend the spectrum of difficulties to adress. Nonetheless, these different challenges did not prevent top participants to reach very high scores, and new dimensions might be explored:

- The table structure and inner-relationships: orientation, concise or nested cells, layout concatenation, multi-valued cells, composed subect (ie subject split into several columns), multi-subjects, hidden subject, etc.
- The out-of-KG-domain data: mentions not present in an existing KG, which is often the case with companies specific data.

It has to be pointed out that the second topic started to be covered in the round 3 of SemTab 2021 where the Schema.org ontology has been used, but the target task was not really consistent with the CTA definition adopted by the community as the ground truth was a mix of types and properties, which may lead to heterogeneous annotations, thus to inconsistent evaluation.

To address the aforementioned dimensions, DAGOBAH team is working on building a hardcore corpus that might be helpful for the community to cope with new challenges.

6 Conclusion

This paper presented DAGOBAH SL 2021 and all the improvements made to cope with the limits of DAGOBAH SL 2020. Through the optimisation of lookup operations and the exploration of richer graph contexts, this new system was able to get very good results during the challenge. Our future work will increase the accuracy for tables with non-explicit or highly ambigous mentions. In particular, token dictionaries (abbrevations, acronyms) were not used up-to-now as we considered it was against the genericity spirit of our system. However, a wider general dictionary built from a huge amount of documents could be a real asset as long as it is generic enough to be used whatever the dataset is. Additionally, when a majority of unmatchable mentions are present in a column, the use of heuristics context scoring strategies are not sufficient. Embeddings-based approches can be leveraged in that case to reduce ambiguities through clustering or document similarity. Pushing this idea a little further, we believe machine-learning approaches based on language models can become a good asset in the most challenging cases. Finally, as our utlimate goal is to address real world data, new challenging corpuses will be used to cope with harder cases.

References

1. Chabot, Y., Labbé, T., Liu, J., Troncy, R.: DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System. In: International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEUR Workshop Proceedings, vol. 2553, pp. 41–48 (2019)
2. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PloS one **10**(6), e0128193 (2015)
3. Huynh, V.P., Liu, J., Chabot, Y., Labbé, T., Monnin, P., Troncy, R.: Dagobah: Enhanced scoring algorithms for scalable annotations of tabular data. In: SemTab@ ISWC. pp. 27–39 (2020)