The LinkingPark System:  
Leveraging the Power of Knowledge Graphs in Tabular Data Productivity

Tables are commonly used to organize information, playing a key role in data analytics, scientific research, and business communication. The ability to automatically extract semantics in tables can empower many downstream applications such as data analytics, robotic process automation (RPA), knowledge base population, etc.

Here we present LinkingPark, an automatic semantic annotation system for tabular data to knowledge graph matching, which enables multiple usage scenarios unlocking extra value from prevalent tabular data. Such annotations cover multiple categories of semantic understanding (see Figure below), like: Cell-Entity Annotation (CEA), Column-Type Annotation (CTA), and Columns-Property Annotation (CPA) altogether.

表格

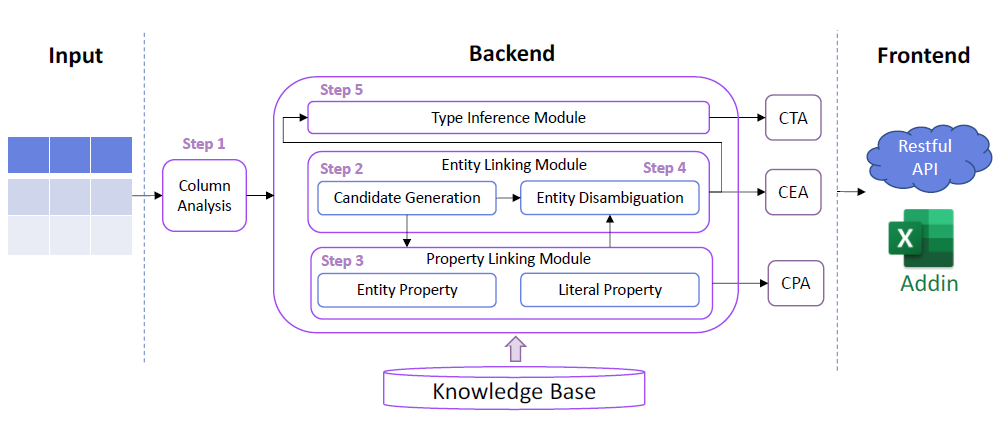
描述已自动生成

LinkingPark features a number of desirable properties including *modular design*, *unsupervised nature*, *stability*, *effectiveness*, *efficiency*, and *flexibility for multilingual support*. Its system backend offers an efficient RESFTful API for programmatic access, as well as an Excel Add-in for easy end-user exploration of data.

**Highlights:**

* Unsupervised nature: LinkingPark adopts an unsupervised approach which makes it easy and cheap to adapt to new knowledge bases, without collecting additional annotated data;
* Stability: LinkingPark does not rely on external services. Its self-contained nature avoids depending on blackbox third-party components whose availability, stability, or efficiency cannot be guaranteed;
* Effectiveness: the system is tailored for real-world scenarios, like handling noisy strings with typos. Its disambiguation algorithm considers multiple factors beyond surface form (e.g., entity popularity information, column-wise type consistency, and row-wise property relatedness). Moreover, LinkingPark builds upon our previous prototype system which won the second prize among 28 teams on the SemTab 2020 competition;
* Efficiency: LinkingPark is designed with the goal of supporting real-time interactions and services, while retaining flexible resource costs. The system can also be configured to fit different deployment or usage requirements along latency, resource allocation/cost, target knowledge base;
* Ease of use: Different scenarios can be enabled by leveraging the offered RESTful API; while a demo interface in the form of an Excel Add-in also allows for direct user interaction with data and annotations for exploration;
* Language-agnostic: By leveraging multilingual resources, LinkingPark can be easily extended to process multilingual tables and link them to a language agnostic knowledge base such as Wikidata. Re-usable components, integration services, REST API, and user-facing interface.

# System Overview



* Modular Architecture
  + Column analysis module
    - Cell data type classification
    - Subject column detection
  + Entity Linking module
    - Candidate generation sub-module
    - Entity disambiguation sub-module
  + Property linking module
    - Entity property linker
    - Literal property linker
  + Type inference module: type voting
  + Knowledge base module: knowledge base storage infrastructure
* Service Interfaces
  + RESTful API
  + Excel Add-in

More detailed information is available on the [website](https://aka.ms/XLKB) (<https://aka.ms/XLKB>) and the available videos and publications.

# Sample Enabled Usage Scenarios

LinkingPark v2 (LPv2) enables multiple new scenarios in tabular data productivity.

1. Data Linking:

Entity/type/property/relationship linking. Data augmentation via type and attribute information.

1. Flash fill:

After a couple examples, auto fill the rest of the missing columns.

图形用户界面, 应用程序, 表格, Excel

描述已自动生成

1. Fact Checking:

Consistency detection/Data cleanup. From simple typos to domain/type/data mismatch and contrasting detected issue against “ground truth”.

表格, Excel

描述已自动生成

1. Auto-complete

During typing, both entity and properties and be automatically completed. Together with Flash fill, this capability can greatly increase productivity.

1. Header suggestion

Help organize data by automatically suggesting standard headers for columns across spreadsheet or databases.

图形用户界面, 应用程序, 表格, Excel

描述已自动生成

# Accessing the System

LPv2 allows service integration via its REST API, as well as direct user access through an add-in to Microsoft Excel. Place contact the Knowledge Computing team as Microsoft Research Asia for access inquiries.