For a survey paper there should be a more comprehensive list of references
The report explains the components of EL and IE well but it fails to give adequate references and explain the various EL techniques in detail including their advantages and disadvantages.
The structure can be improved, eg the introduction is too long and already covers things like task definition which should be in later sections
the difference or relation of REFINED to other methods can be made clearer

Exploring the Core Methodologies and Innovative Role of ReFinED in Entity Linking: A Comprehensive Survey

Yu-Chuan Cheng
Scientific Computing, Heidelberg University
yu-chuan.cheng@stud.uni-heidelberg.de

Abstract

According to (Shen et al., 2015), my survey of ReFinED in entity linking explores its core methodologies across key modules: Candidate Entity Generation, Ranking, and Unlinkable Mention Prediction. We delve into entity linking dimensions, including applications, features, metrics, and relevant datasets. This perspective highlights ReFinED's innovative role in entity linking.

1 Introduction

1.1 Motivation

In today's era of digital transformation, effectively harnessing unstructured textual data is paramount. Advanced language models like OpenAI's ChatGPT and Anthropic's models exemplify this trend, revolutionizing data interaction. Central to these advances is Entity Linking (EL), a core task in Natural Language Processing (NLP). EL connects unstructured text with structured entities, a task pivotal to various applications. Despite its apparent simplicity, EL poses challenges due to language ambiguity, expansive entity scope, and data source disparities. EL's significance is manifest in streamlining workflows. Law firms utilize EL to link case-specific terms to legal references, while businesses gain searchable insights from historical data. EL also enhances Customer Relationship Management (CRM) systems for improved customer engagement.

This paper surveys EL approaches, dissecting methods, strengths, limits, and real-world roles. Pioneering ReFinED(Ayoola et al., 2022), a novel EL method, is explored, showcasing its benchmark-leading performance. This survey guides newcomers and researchers, clarifying EL's landscape amidst rising NLP adoption. It underscores ReFinED's role, delving into its innovative model components. With EL's increasing relevance in data-driven tasks, our survey contributes insights into methods, applications, and induction, notably ReFinED's excellence in EL.

1.2 Task Description

Before delving into the intricacies of Entity Linking (EL), grasp these foundational notions: entity mentions and knowledge bases, pivotal in EL's domain. In Natural Language Processing (NLP), an entity mention denotes a named entity occurrence in text. These can span people, organizations, dates, and more. For example, "Apple Inc." and "Steve Jobs" in "Apple Inc. was founded by Steve Jobs."

A knowledge base serves as a structured repository, housing facts about entities and their relationships. Examples like Wikipedia and Wikidata abound. Entities are nodes, each with unique IDs and attributes, interconnected by relationships like "founded by." Entity Linking binds text mentions to distinct knowledge base entities. In our example, EL matches "Apple Inc." and "Steve Jobs" to their knowledge base counterparts. EL's complexity stems from language ambiguity. "Apple" might denote multiple entities, demanding context comprehension for accurate linking. EL bridges unstructured text and structured data, enabling data utilization.

A typical entity linking system is composed of three essential modules, each playing a vital role in linking entity mentions from unstructured text to their corresponding entities in a knowledge base:

1. **Candidate Entity Generation:** In this crucial module, the entity linking system focuses on processing each entity mention, denoted as "m" from a set "M," and aims to filter out irrelevant entities from the knowledge base. The goal is to retrieve a candidate entity set that contains potential entities to which the mention "m" may refer. To achieve this, state-of-the-art entity linking systems employ a diverse array of techniques.

Some common techniques used in this module include name dictionary-based methods, where pre-defined dictionaries help identify candidate entities. Surface form expansion, on the other hand, involves expanding the context around the mention in the local document to enrich the candidate entity set. Additionally, methods based on search engines are utilized to extract candidates based on web search results.

2. **Candidate Entity Ranking:** Once the candidate entity set is assembled, researchers face the challenge of ranking these candidates to identify the most likely link for the mention "m." This ranking is crucial as the candidate entity set often contains more than one potential entity.

   Researchers leverage different types of evidence to assess the relevance of each candidate entity to the mention "m." Various features and contextual information are considered to determine the entity that best aligns with the mention. However, ReFinED involved fine-grained entity typing which is a part of Candidate Entity Ranking.

3. **Unlinkable Mention Prediction:** Dealing with the issue of predicting unlinkable mentions is another essential aspect of entity linking. In some cases, the top-ranked entity identified in the Candidate Entity Ranking module might not be the correct target entity for the mention "m."

   To address this challenge, certain entity linking systems incorporate an unlinkable mention prediction module. This module validates whether the top-ranked entity is indeed the correct link for the mention. If not, the system assigns a special label "NIL," indicating that the mention "m" cannot be linked to any entity in the knowledge base.

   In the ReFinED paper, a distinctive candidate is introduced for the NIL entity, assigned an un-normalized score of 0, denoting the absence of correct candidate entities. This highlights ReFinED's incorporation of Unlinkable Mention Prediction, selecting either a Knowledge base entity or NIL with the highest combined score during entity linking (EL) inference. Thus, ReFinED encompasses multiple loss functions in its approach, enhancing its versatility and accuracy.

1.3 Applications

As outlined in the introductory section, entity linking serves as a vital component in numerous functions. In this section, we explore a variety of its common applications.(Shen et al., 2015)

1. **Information Extraction:**Entity linking is fundamental in extracting structured information from unstructured text. It helps identify and resolve the entities mentioned in the text(Lin and Etzioni, 2012), providing a crucial step towards understanding the information contained within it. This understanding can then be used in numerous applications, such as building knowledge graphs, automated report generation, and more.

2. **Question Answering Systems:**In systems like Siri, Alexa, or Google Assistant, entity linking is crucial to understand the user's query accurately and provide a relevant answer. For instance, if a user asks "What movies has Brad Pitt starred in?", the system must identify "Brad Pitt" as an entity and link it to the corresponding entity in a knowledge base to retrieve the list of movies.

3. **Search Engines:**Google and other search engines use entity linking to enhance their search results. For example, if you search "Apple founder", Google recognizes "Apple" as the tech company and "founder" as a property related to it, and returns "Steve Jobs" and "Steve Wozniak" as results. This kind of intelligent search capability is made possible by entity linking.

4. **Text Summarization:**Entity linking aids in the summarization of large amounts of text by identifying key entities and their relations. By linking entities to a knowledge base, a summarization system can better understand the text's context and generate more accurate and meaningful summaries.

5. **Semantic Web and Knowledge Graph Construction:**Entity linking plays a critical role in building knowledge graphs, where nodes represent entities and edges represent relationships between them. By accurately linking entities, we can establish connections between

different pieces of information, thereby enabling the construction of the semantic web.

6. **Named Entity Disambiguation:**Entity linking is used to resolve the ambiguity of named entities in text. For example, a system can use entity linking to determine whether "Apple" refers to the tech company, the fruit, or something else, based on the surrounding context.

7. **Cross-Lingual and Multilingual Applications:**In applications involving multiple languages, entity linking can help map an entity mentioned in one language to its equivalent in another. For instance, a news article about "Barack Obama" in French can be linked to the corresponding English entity in a knowledge base.

Entity linking, thus, serves as a cornerstone in various NLP tasks and applications, enhancing their capacity to understand, interpret, and generate meaningful output from textual data.

1.4 Preliminaries

A crucial component of the entity linking task is the knowledge base, which serves as a structured repository of information about the world's entities. Within a knowledge base, one can find details about specific entities such as "Julius Robert Oppenheimer" and "Oppenheimer(film)," along with their respective semantic categories like "Scientist" and "City." Additionally, knowledge bases capture the relationships that exist between entities, exemplified by associations like "bornIn" between "Julius Robert Oppenheimer" and "Oppenheimer."

movie?

In the following section, we provide a concise introduction to three widely utilized knowledge bases and one state-of-the-art EL model in the field of entity linking. These knowledge bases play a pivotal role in enabling the linking of entity mentions within unstructured text to their corresponding entries in the knowledge base, thus facilitating the extraction of structured information from unstructured data.

The reference does not seem to be about wikipedia in general

1.4.1 Wikipedia

[1] Wikipedia(Yano and Kang, 2016) stands as a remarkable online encyclopedic resource, shaped collaboratively by thousands of volunteers worldwide. As a free and multilingual platform, it has grown into the largest and most widely accessed Internet encyclopedia today, continually expanding with new content. The core unit of information on Wikipedia is an "article," each uniquely identified and dedicated to defining and describing specific entities or topics.

With over 4.4 million articles in the English edition alone, Wikipedia boasts extensive coverage of named entities, encompassing a vast repository of knowledge about notable individuals, organizations, locations, and more. This wealth of information renders Wikipedia a valuable resource for various natural language processing tasks, particularly entity linking.

The structure of Wikipedia offers a range of essential features for entity linking purposes, including dedicated "entity pages," precise "article categories," helpful "redirect pages" that assist in resolving mentions to their corresponding entities, and "disambiguation pages" to handle potentially ambiguous references. Moreover, the interconnectivity of Wikipedia articles through hyperlinks further enhances the ability to navigate and link relevant entities within the vast ecosystem of Wikipedia's knowledge.

Due to its richness, comprehensiveness, and user-driven nature, Wikipedia serves as an indispensable source for advancing research and applications in entity linking and other language-related tasks, contributing significantly to the understanding and organization of knowledge in the digital age.

1.4.2 Wikidata

The key objective of Wikidata(Krötzsch, 2014) is to provide a structured and machine-readable representation of factual knowledge about the world. Wikidata serves as a central hub of information, housing data on various entities like individuals, locations, organizations, events, and abstract concepts. Each entity within Wikidata is assigned a unique identifier, and the data pertaining to these entities is stored in the form of structured statements known as "claims." Wikidata currently contains 105,381,089 items. 1,938,944,298 edits have been made since the project launch.(untill 24.07.2023)

Key features of Wikidata include:

- **Structured Data:** Unlike Wikipedia, which primarily consists of unstructured text in articles, Wikidata stores data in a structured format using statements and properties. These properties define specific attributes of an en-

[1] http://www.wikipedia.org/

tity and their corresponding values.

- **Multilingual:** Wikidata is multilingual, meaning it supports data in multiple languages. Entities and their attributes can be described in various languages, making it a valuable resource for cross-lingual applications.
- **Interlinking with Wikipedia:** Wikidata is closely linked with Wikipedia, enabling the integration of structured data with Wikipedia articles. This connection allows information to be displayed across different language versions of Wikipedia.
- **Data Queries:** Wikidata provides a powerful query service that allows users to retrieve specific information using the SPARQL query language. This feature enables researchers and developers to access data in a flexible and targeted manner.
- **Open Collaboration:** Like other Wikimedia projects, Wikidata is an open collaborative platform. Anyone can contribute to and edit the data, making it a constantly evolving and up-to-date knowledge base.

Overall, Wikidata is an essential and constantly expanding knowledge base that has a significant impact on storing, sharing, and utilizing structured data for various purposes in the digital era. Its collaborative nature and structured methodology make it a valuable asset for researchers, developers, and the general public alike.

1.4.3 DBpedia

DBpedia (Auer et al., 2007) stands as a multilingual knowledge base sourced from Wikipedia, extracting structured data like infobox templates, categorization details, geo-coordinates, and external webpage links. The English iteration of DBpedia encompasses 4 million entities, with 3.22 million conforming to a unified ontology. Furthermore, it seamlessly updates in tandem with Wikipedia's alterations.

1.4.4 BLINK

(Wu et al., 2020), a cutting-edge Entity Linking model introduced by Facebook, has revolutionized the field by leveraging Wikipedia as its target knowledge base. The process of linking entities to their corresponding Wikipedia entries is commonly referred to as "Wikification."

At its core, BLINK utilizes a two-stage approach, hinging on the power of fine-tuned BERT (Bidirectional Encoder Representations from Transformers) architectures. This unique strategy sets BLINK apart as an advanced and efficient entity linking solution.

Stage 1:Dense Space Retrieval. In the first stage, BLINK employs a bi-encoder to independently embed the context of the entity mention and the descriptions of candidate entities. This creates a dense space representation, which facilitates efficient retrieval of relevant candidates for each mention.

Stage 2:Comprehensive Examination. Having narrowed down potential candidates in Stage 1, BLINK proceeds to conduct a more in-depth analysis using a cross-encoder. This component concatenates the mention and entity text, capturing their interrelations more comprehensively.

BLINK's ingenuity and sophistication translate to remarkable performance on multiple datasets. It surpasses the benchmark, achieving state-of-the-art results in the domain of entity linking.

1.5 Outline

the introduction section is too long. This overview should be on page 1 or 2, not page 4

In this survey, we thoroughly review and analyze the techniques used in the ReFinED system for entity linking. We attribute these techniques to the relevant modules, such as Candidate Entity Generation, Candidate Entity Ranking and Unlinkable Mention Prediction. We also explore features and evaluation methods. Section 2 presents the features for these modules. Section 3 covers the evaluation of entity linking systems. In Section 4, we conclude and discuss ReFinED's contributions and potential future research directions.

2 Methodologies

2.1 Candidate Entity Generation

Candidate entity generation approaches primarily rely on string comparison between the surface form of the mentioned entity and the names of entities in a knowledge base. In this section, we review the main approaches used for generating candidate entity sets for entity mentions.

ReFinED is mainly involved in the first and second sections of candidate entity generation. Specifically, in Section 2.1.1, we describe the name dictionary-based techniques. These involve matching entity mentions with pre-defined dictionaries of known entity names.

In Section 2.1.2, we explore the approaches based on search engines. These techniques leverage web search engines to retrieve candidate entities related to the mentioned entity.

By understanding these key approaches, we gain insights into ReFinED's strategies for generating candidate entities and effectively linking entity mentions to their corresponding entities in knowledge bases.

2.1.1 Name Dictionary Based Techniques

In the context of Wikipedia-based entity linking systems, the structure of Wikipedia offers valuable features to generate candidate entities. These features include entity pages, redirect pages, disambiguation pages, bold phrases from the first paragraphs, and hyperlinks in Wikipedia articles.

To construct an offline name dictionary for candidate entity generation, these systems leverage various combinations of these features(Guo et al., 2013). The name dictionary contains an extensive collection of information about different names associated with named entities, encompassing name variations, abbreviations, confusable names, spelling variations, nicknames, and more.

The name dictionary takes the form of a key-value mapping. The key column consists of a list of names, while the value column contains sets of named entities that can be associated with each respective name. In the case of ReFinED, we specifically focus on its involvement in constructing the dictionary using features extracted from Wikipedia.

The process of building the dictionary involves carefully utilizing Wikipedia features, enabling efficient and comprehensive candidate entity generation. These techniques, when integrated into entity linking systems like ReFinED, contribute to robust and accurate entity linking, leveraging the wealth of information present in Wikipedia's vast repository of knowledge.

The dictionary is constructed by leveraging features from Wikipedia as follows:

- **Entity pages:**In Wikipedia, each entity page focuses on describing a single entity and is titled with the most common name for that entity, such as "Apple" for the technology company. In the Dictionary, the title of the entity page (name "key") is linked to the entity described on that page (value). This facilitates efficient entity linking by connecting entity mentions to their corresponding entities in the knowledge base.
- **Redirect pages:**In the Wikipedia-based entity linking system, redirect pages play a crucial role. These pages exist for alternative names that can refer to an existing entity in Wikipedia. For instance, the article "Apple Inc." has a redirect page pointing to the entity "Apple," which represents the same entity. In the entity linking process, the title of the redirect page, such as "Apple Inc.," is added to the key column in the Dictionary as a name "key," and the pointed entity, "Apple," is added as the corresponding value. This allows efficient linking of entity mentions to their corresponding entities in the knowledge base, incorporating synonym terms, abbreviations, and other variations of entities for comprehensive candidate entity generation.
- **Disambiguation pages:**Disambiguation pages in Wikipedia serve to separate multiple entities with the same name. They contain references to those entities, helping to resolve ambiguity. For instance, the disambiguation page for "Nikola" lists 12 associated entities, including the vehicle company and the German TV series. In the Dictionary, the title of each disambiguation page (name "key") is linked to the entities listed on that page (value). This allows efficient extraction of abbreviations and aliases of entities, aiding in accurate entity linking by associating entity mentions with their appropriate entities in the knowledge base.
- **Hyperlinks in Wikipedia articles:**Hyperlinks in Wikipedia articles provide valuable information for entity linking. They link to pages of the entities mentioned in the article and often contain anchor text, which serves as synonyms and name variations of the linked entity. In the Dictionary, the anchor text of the hyperlink (name "key") is linked to the entity page it points to (value). This enables the extraction of useful synonyms and variations, contributing to accurate entity linking by connecting entity mentions with their corresponding entities in the knowledge base.

Entity linking systems utilize the features described earlier from Wikipedia to construct a dictionary. Additionally, some studies explore query

click logs and web documents to find entity synonyms, enhancing the dictionary's scope.

With the constructed dictionary, the simplest approach for generating candidate entity sets involves exact matching between the name keys in the key column and the entity mentions. If there is a match between a name key and an entity mention, the corresponding set of entities in the value column is added to the candidate entity set. This approach facilitates efficient and accurate linking of entity mentions to their potential entities in the knowledge base.

The ReFinED model's loss computation incorporates the features mentioned earlier, including Entity description and Mention representation. However, Entity typing and Entity prior (based on Entity popularity) are part of the Candidate Entity Ranking module. Together, these components enable ReFinED to effectively compute scores or loss and pick candidate entities, enhancing the entity linking performance by leveraging a comprehensive set of features.

lower case

2.1.2 Methods Based on Search Engines

Some entity linking systems take advantage of web search engines, such as Google, to identify candidate entities by leveraging information from the entire web. Additionally, the Wikipedia search engine is also utilized to retrieve candidate entities. By querying the Wikipedia search engine based on keyword matching, relevant Wikipedia entity pages are returned, enabling the system to build a list of potential candidate entities for entity linking. These search engine-based approaches enhance the entity linking process by tapping into a vast array of information available on the web and in Wikipedia, thus expanding the scope of candidate entity generation.

2.1.3 Surface Form Expansion from the LocalDocument

Some entity linking systems employ surface form expansion techniques to identify possible expanded variations of entity mentions, such as full names or acronyms. These expansions are obtained from the associated document where the entity mention appears. Subsequently, these expanded forms are leveraged to generate the candidate entity set using methods like the name dictionary-based techniques described earlier.

Surface form expansion techniques fall into two categories: heuristic-based methods and supervised learning methods. However, it's important to note that ReFinED does not initially build its own dictionary. Instead, it uses Wikipedia as a pre-training dataset and fine-tunes with the AIDA-CoNLL dataset(Hoffart et al., 2011). Expansion techniques are applied during the Wikipedia dataset building process rather than after its construction. These techniques play a crucial role in ensuring comprehensive and accurate candidate entity generation from the vast pool of entity mentions found in the Wikipedia dataset.

2.2 Candidate Entity Ranking

In the previous section, we discussed methods for generating a candidate entity set for each entity mention. In many cases, this set contains more than one candidate entity, with an average of 73 candidates on the CoNLL dataset used in (Tjong et al., 2003), as mentioned in the ReFinED paper. The next step is to utilize various types of evidence to rank these candidate entities and select the most appropriate entity as the mapping entity for the entity mention. This process is achieved through the Candidate Entity Ranking module, a crucial component in the entity linking system. We can categorize candidate entity ranking methods into two main types:

1. **Supervised ranking methods:** These approaches rely on annotated training data to learn how to rank the candidate entities in the entity set. Some examples of supervised ranking methods include:
    - **Binary classification methods:** These models classify candidate entities as relevant or irrelevant to the entity mention.
    - **Learning to rank methods:**These methods aim to directly optimize the ranking of candidate entities using various machine learning techniques.
    - **Probabilistic methods:** These approaches leverage probabilistic models to estimate the likelihood of candidate entities being the correct mapping for the entity mention.
    - **Graph-based approaches:** These methods construct graphs that represent relationships between entities and use graph-based algorithms to rank the candidates.
2. **Unsupervised ranking methods:** Unlike supervised methods, these approaches do not rely on manually annotated data for training.

Instead, they use unlabeled corpora to rank candidate entities. Some examples of unsupervised ranking methods include:

- **VSM (vector space model) based methods:**These are a type of unsupervised ranking methods used for entity linking. The candidate entity with the highest similarity score is selected as the mapping entity for the entity mention. VSM based methods differ in their methods of vectorial representation and vector similarity calculation.
- **information retrieval-based methods:**These methods are utilized in entity linking and operate without supervision. They involve selecting the candidate entity with the highest similarity score as the mapped entity for a given entity mention. Information retrieval-based techniques view candidate entity ranking as an information retrieval problem, treating each candidate entity as a distinct document. For each entity mention, a search query is formulated, and similarity scores are computed. The KL-divergence is one of the metrics used for measuring this similarity.

2.2.1 Features

This section examines valuable features for candidate entity ranking. These features are categorized as context-independent (relying on entity info) and context-dependent (linked to entity's context). The latter encompasses both textual context and other linked entity mentions within the document.

similar feedback: you list relevant features and methods but need to give references and explain a bit more on their details and strength

Context-Independent Features

- **Name string comparison:** The name string comparison between the entity mention and the candidate entity is the most direct feature that one may use. Many string similarity measures have been used in the name comparison, including edit distance, Dice coefficient score, character Dice, skip bigram Dice, and left and right Hamming distance scores.
- **Entity Popularity:** Another context-independent feature of significance in entity linking involves the popularity of a candidate entity in relation to the mentioned entity. This feature provides the prior likelihood of a candidate entity appearing alongside the entity mention. Notably, it's (it is) observed that candidates within the same entity mention category possess varying popularity levels. Some entities are notably uncommon or obscure in relation to a given mention. For instance, considering the entity mention "Oppenheimer" ~~(film),~~ the candidate entity "ulius Robert Oppenheimer" a" holds a lower occurrence rate compared to "Oppenheimer" (film). Frequently, the mention "Oppenheimer" pertains to the film rather than the individual sharing the same name.
- **Entity Type:** This attribute signifies the alignment between the entity mention's type (such as people, location, and organization) in the text and the corresponding candidate entity type within the knowledge base. In cases where the type is unidentified in the knowledge base, consultation with DBpedia is conducted.

While context-independent features offer value, they solely draw upon information from the entity mention and the candidate entity. Yet, incorporating context-specific features, grounded in the entity mention's surroundings, is crucial. However, it's noteworthy that the ReFinED model doesn't (does not) encompass context-dependent features, and a comprehensive discussion of this aspect is beyond the scope here.

Context-Dependent Features

- **Textual Context:** A fundamental feature concerning textual context involves assessing the textual resemblance between the context surrounding the entity mention and the document linked to the candidate entity. Various representation formulations allow conversion of these contextual texts into vectors. Diverse techniques, such as dot-product, cosine similarity, Dice coefficient, word overlap, KL divergence, n-gram based metrics, and Jaccard similarity, have been applied to gauge vector similarity.
- **Coherence Among Mapped Entities:** The significance of textual context surrounding an entity mention in entity linking is undeniable. Furthermore, when considering entity mentions within a document, the linkage of other related entity mentions holds relevance. Modern entity linking systems often assume that

a document predominantly pertains to coherent entities within a shared theme or related topics. This topical coherence is harnessed to collectively link entity mentions in the document. Consequently, leveraging the topical coherence feature among mapped entities in a document enhances the process of entity linking.

2.3 Unlinkable Mention Prediction

In the previous section, we delved into techniques for ranking candidate entities. However, real-world scenarios often involve entity mentions lacking knowledge base matches, necessitating strategies for unlinkable mention prediction. Various approaches address this challenge.

Some studies assume comprehensive knowledge bases and omit unlinkable mention prediction. Simple heuristic methods label mentions with no candidates as unlinkable (NIL). Many entity linking systems employ supervised machine learning, using binary classification for unlinkable mention prediction. Certain approaches integrate unlinkable prediction into entity ranking. A distinct NIL entity is added to candidate sets. If NIL ranks first, the mention is unlinkable; otherwise, the top-ranked entity is selected. An advanced probabilistic model handles unlinkable mention prediction seamlessly. It introduces a NIL entity, comparing probabilities to identify unlinkable mentions.

3 Evaluation

The assessment of entity linking systems is usually per-formed in terms of evaluation measures, such as precision, recall, F1-measure, and accuracy.

$$recall = \frac{\text{correctly linked entity mentions}}{\text{entity mentions that should be linked}}$$

$$precision = \frac{\text{correctly linked entity mentions}}{\text{linked mentions generated by system}}$$

$$F_1 = \frac{2 \times precision \times recall}{\text{precision+recall}}$$

However, there are two different F1 metric:

1. **Macro F1 Score: Striving for Equitable Evaluation(each class)**
2. **Micro F1 Score: Emphasizing Global Performance** In contrast, the micro F1 score prioritizes the big picture. It aggregates true positives, false positives, and false negatives across all classes before calculating the F1 score. This approach emphasizes the overall classification performance of the model, making it a suitable choice when the primary concern is the model's ability to correctly classify instances as a whole. Moreover, the micro F1 score's equal weighting of instances makes it robust in scenarios where imbalanced datasets are the norm.

4 Conclusion

ReFinED, a supervised learning model, harnesses Wikidata for pretraining and AIDA-CONLL for fine-tuning. Wikidata's rich entity typing, spanning 90 million entities, underpins ReFinED's Probabilistic approach. Optimizing entity candidates through context-dependent features, entity typing, and popularity, ReFinED achieves performance with 231 forward passes.

Comparing BLINK and ReFinED, both use cross-entropy as loss, but their inference speeds differ due to architecture variations. ReFinED excels in balancing data features and accuracy, urging reflection on performance versus precision trade-offs. Potential lies in dataset features, although implementation methods are elusive.

This paper surveys ReFinED in entity linking. It examines core methodologies, from Candidate Entity Generation to Unlinkable Mention Prediction, and probes applications, features, and metrics. A panoramic view elucidates ReFinED's innovative role.

References incomplete, eg missing journal or coference name

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. Refined: An efficient zero-shot-capable approach to end-to-end entity linking.

Stephen Guo, Ming-Wei Chang, and Emre Kıcıman. 2013. To link or not to link? a study on end-to-end tweet entity linking.

Johannes Hoffart, Mohamed Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol,

Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text.

Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base.

Thomas Lin and Oren Etzioni. 2012. Entity linking at web scale.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27:443–460.

Erik Tjong, Kim Sang, and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval.

Tae Yano and Moonyoung Kang. 2016. Taking advantage of wikipedia in natural language processing.