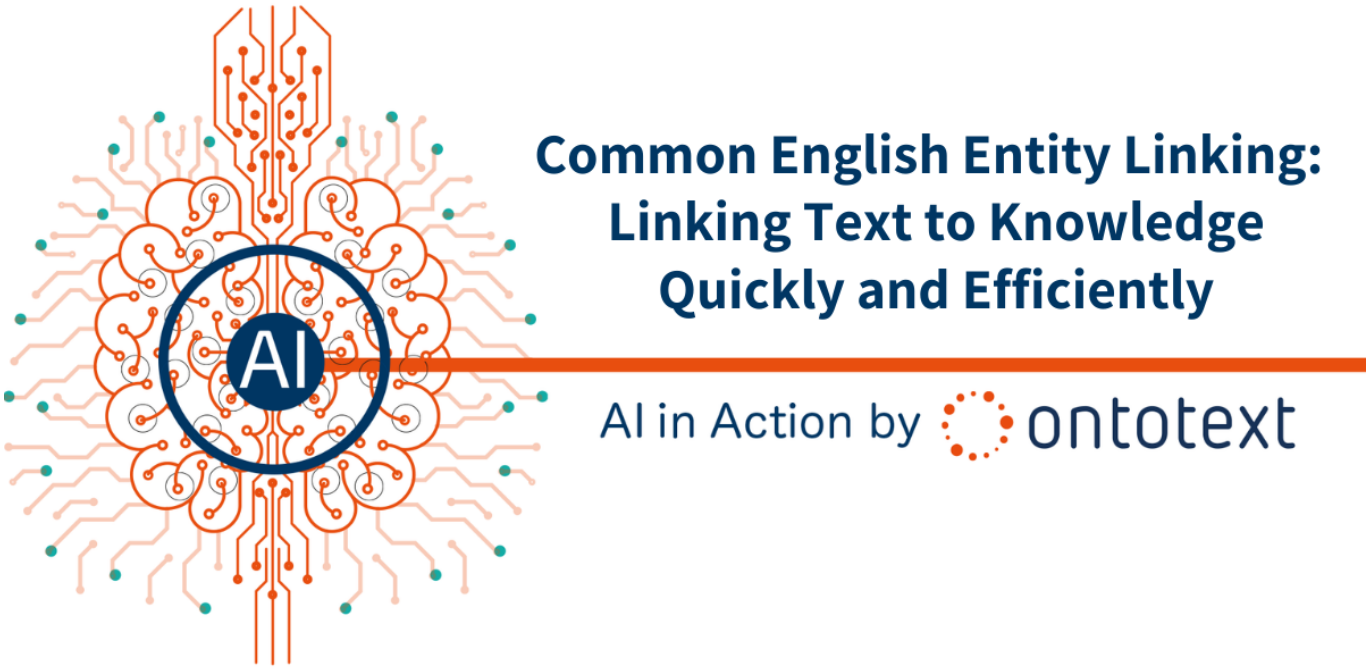


This is part of Ontotext's AI-in-Action initiative aimed at enabling data scientists and engineers to benefit from the AI capabilities of our products.

Entity linking is the process of automatically linking entity mentions from text to the corresponding entries in a knowledge base. It has been an important capability for Ontotext ever since we dove into Natural Language Processing (NLP), as it is a crucial aspect of the interplay between text analysis and knowledge graphs.

Usually, entity linking consists of two main tasks – named entity recognition and entity disambiguation. Named entity recognition involves recognizing mentions of entities in text, while entity disambiguation assigns a knowledge base entity to each mention. We've had a fair share of such implementations over the years. One such example is the tagging behind News On the Web. It is a demonstrator designed to showcase a field in which Ontotext has built a number of solutions applying semantic technology to enhance publishing content and editorial workflows.

Introducing Ontotext Common English Entity Linking

Now comes our next-generation offering – Common English Entity Linking (CEEL). CEEL is an end-to-end transformer-based model that includes mention detection and uses entity types

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

and entity descriptions to perform entity linking. Out of the box, the model links to Wikidata IDs. It is also flexible enough to support adaptation to other large-scale knowledge bases and perform zero-shot entity linking. The combination of speed, accuracy and scale makes CEEL an effective and cost-efficient system for extracting entities, even when talking about web-scale datasets.
CEEL is now immediately available as part of our text analysis offerings, coming preconfigured as part of the new version of the Ontotext Metadata Studio (OMDS). You can also use it as a service to integrate in your own products or we can feature it as part of a broader custom solution we build for you. Documents loaded in GraphDB can be sent for processing to CEEL with a simple SPARQL query – the resulting tags will be loaded directly in the graph.

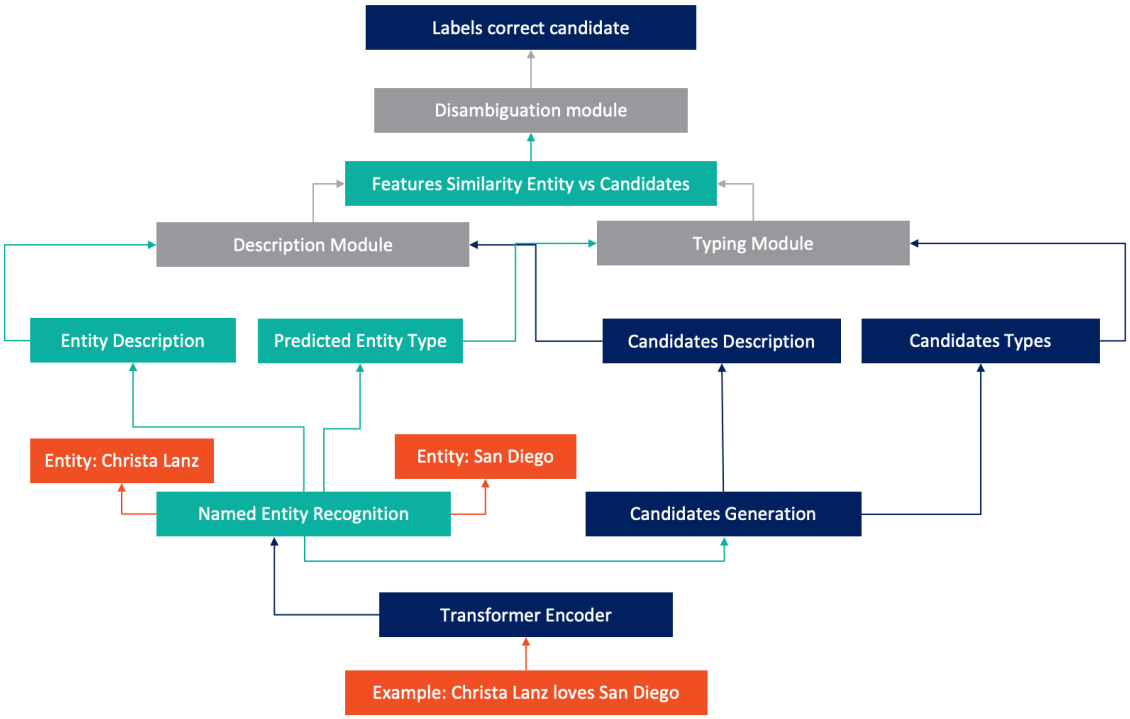
Problem

Some of the current popular entity linking systems, such as GENRE, BLINK, show great performance on standard datasets, but have several limitations when used in real-world applications.
First of all, they are computationally heavy, which makes large-scale processing expensive. Also, most entity linking systems are designed to link to specific knowledge bases (typically Wikipedia) and cannot be easily adapted to other knowledge bases. Another important point is that existing methods cannot link text to entities that were introduced to the knowledge base after training (a task known as zero-shot entity linking). In other words, they must be frequently retrained to be kept up-to-date, which requires the expensive process of gold corpus preparation.
CEEL addresses all these problems. It is efficient, it provides great accuracy and, due to its zero-shot capability, you do not need to retrain it when linking to other databases.

Architecture

Entity linking can be a challenging task because entity names are often ambiguous. That's why such models must use context in order to differentiate between entities, just as humans do. Accounting for the context around entities is best achieved with the use of transformer-based models and that's what CEEL is based on.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



The diagram above illustrates how CEEL's architecture uses entity types and entity descriptions of generated candidates for each recognized mention to perform entity linking. The initial statement is "Christa Lanz loves San Diego". Then the end-to-end system identifies the entities in the text: "Christa Lanz" and "San Diego", via the named entity recognition module. Next, the system disambiguates the correct candidate through the entity disambiguation component.

Named Entity Recognition

This is a token classification task and the tokens encoded in the input text document use a pre-trained transformer model. Then tokens are classified based on the contextualized token embeddings.

Entity Disambiguation

The entity disambiguation stage consists of the following components:

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

- **Candidate generation:** this is a standard step where, based on gazetteers, a set of candidates are proposed for all recognized entities from the named entity recognition stage.
- **Entity typing:** the entity typing module estimates how similar the predicted entity types are for the recognized mention and the real entity types for the generated candidates.
- **Entity description:** this shows how relevant the text description of the proposed candidates is to the recognized mentions. The entity description module resembles a bi-encoder architecture. It consists of two separate transformer encoders — one for encoding the generated candidate description and another for encoding recognized mentions. These encoders produce embeddings for the mention and the candidates and compare how similar they are.
- **A combination of the above:** CEEL computes a combined score by concatenation of the features for the entity typing score, entity description score for the candidates and passing them through a set of neural network layers.

All the components of the entity disambiguation module were trained on the Wikipedia corpus for a few epochs and fine-tuned on AIDA CoNLL-YAGO.
This approach has certain benefits. First of all, although it is transformer-based, it is very efficient on CPU's, as the architecture is heavily optimized and inference through the different modules happens in one forward pass. Also, it has zero-shot functionality, meaning that it is not necessary to re-train or fine-tune the model for different corpora. In addition, the model links to Wikidata entities.

Evaluation

We have evaluated CEEL's performance against several of the most commonly used benchmarks for entity linking. They show its competitiveness against other alternatives such as Google NLP and Facebook's MGENRE. Aside from showing good performance results, the system is about 10 times faster than other competitive services outlined here. For example, for a corpus with an average length of 1100 characters per document CEEL vs Facebook GENRE and MGENRE mean inference time on CPU is 5s vs 52s in favor of CEEL. This makes it an effective and cost-efficient system for extracting entities from large-scale data collections.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

	AIDA/CONLL-YAGO*		MSNBC		TWEEKI_GOLD		AQUAINT	
	Entity Recognition	Entity Linking	Entity Recognition	Entity Linking	Entity Recognition	Entity Linking	Entity Recognition	Entity Linking
Ontotext CEEL	96%	76%	91%	66%	87%	60%	N/A	45%
Google NLP	79%	58%	84%	69%	78%	65%	N/A	34%
Facebook GENRE	77%	61%	44%	31%	69%	51%	N/A	28%
mGENRE**	65%	53%	33%	18%	72%	57%	N/A	27%
Ontotext Tag	N/A	62%	N/A	N/A	N/A	N/A	N/A	N/A

Source: Benchmarks evaluation is performed internally. N/A: Either dataset has no offset annotations or results are not available.
** CEEL is fine-tuned on AIDA/CONLL-YAGO Train dataset; evaluation is performed on Test B.*

*** Since mGENRE is only an Entity Disambiguation model, the Entity Recognition component was developed by Ontotext.*

Use Cases

CEEL holds immense potential for various use cases across industries.

- **In Media and Publishing**, CEEL can be employed to enhance content discoverability by linking mentions in news articles to relevant knowledge base entries. In this way, it will provide readers with instant access to additional context.
- **In the Business Intelligence sector**, it can streamline information extraction from large-scale datasets such as RSS feeds. This will enable organizations to quickly analyze and comprehend market trends.
- **In academic research,** CEEL can aid in the automated extraction and categorization of entities from scholarly articles, facilitating more efficient literature reviews and knowledge synthesis.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Availability

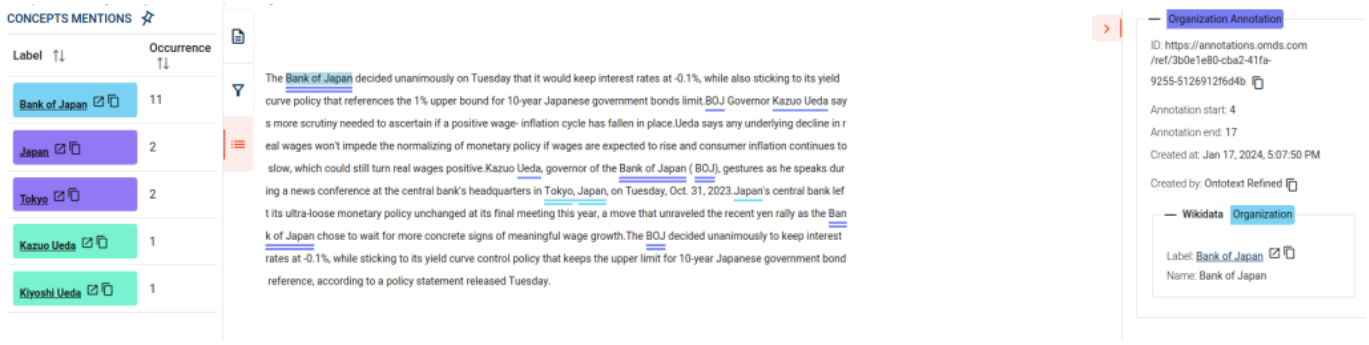
CEEL can be integrated into any applications, workflows, or systems where you need to extract entities from text and link them to factual knowledge. Because of the efficient architecture, CEEL's inference is in the range of a few seconds per document on a CPU hardware. This makes it cost-effective even when you need to process large corpora.
CEEL can be made available in a wide range of options – from the technical to the user-friendly. Directly through GraphDB, it can be called via the text mining plug-in and the output can be serialized in whatever form you prefer. This could be done with a simple SPARQL query, such as this one:

```
PREFIX : <https://www.ontotext.com/textmining#>
PREFIX inst: <https://www.ontotext.com/textmining/instance#>

INSERT DATA {
inst:ceelService :connect :Ces;
:service "" ;
:header "X-API-Key: ";
:header "Accept: application/vnd.ontotext.ces+json";
:header "Content-type: text/plain".
```

How to use such a service to handle the information extracted from it and materialize it as annotations in the graph is described in this section of the GraphDB documentation.
CEEL can also be provided as a service exposed through a programming interface to be used by your user interfaces, products or internal systems. Furthermore, as a part of Ontotext's offerings, CEEL can also be integrated in our products and the custom systems we build.
Finally, a user-friendly way to use CEEL out of the box would be through an integration in OMDS. In this way, you can also evaluate its performance against a benchmark you set and orchestrate its output around additional NLP services – either developed by you, or provided by us or third parties.
The easiest way to use CEEL is by tagging documents in Ontotext Metadata Studio with it. Here follows an example from news content with references to Wikidata concepts highlighted and available for review and curation:

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



When processing a document, CEEL will recognize the mentions, along with the corresponding inline annotations in the text and will link them to the respective Wikidata entity IDs.
And the cherry on top: CEEL can be mapped to the specific set of entities and relationships from your domain of interest. Its architecture allows for flexibility and CEEL can easily be adopted to suit your domain and knowledge base instances. Since the knowledge base candidate generation step is decoupled from the model itself, through lightweight customization, we can quickly transform your situation from no text analysis capabilities to full coverage working with the entities you care about.

To Sum it Up

Introducing Ontotext's CEEL represents a leap forward in what we can offer in the field of entity linking, addressing key challenges faced by existing systems. Its efficiency, accuracy and scalability make it a valuable asset for organizations dealing with vast amounts of textual data.
The transformer-based architecture, optimized for CPU usage, ensures swift processing without compromising computational resources. The zero-shot functionality eliminates the need for frequent retraining and provides a cost-effective way to adapt it for real-world applications by linking seamlessly to different knowledge bases.

Want speedy and accurate large-scale data extraction with unparalleled efficiency?

Contact Us or Try it in GraphDB

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com