

Neural Entity Linking: A Survey of Models based on Deep Learning

Özge Sevgili^{1*}, Artem Shelmanov^{2*}, Mikhail Arkhipov³,
Alexander Panchenko², and Chris Biemann¹

¹Universität Hamburg, Hamburg, Germany

{sevgili, biemann}@informatik.uni-hamburg.de

²Skolkovo Institute of Science and Technology, Moscow, Russia

{a.shelmanov, a.panchenko}@skoltech.ru

³Moscow Institute of Physics and Technology, Dolgoprudny, Russia
arkhipov@yahoo.com

Abstract

In this survey, we provide a comprehensive description of recent neural entity linking (EL) systems. We distill their generic architecture that includes candidate generation, entity ranking, and unlinkable mention prediction components. For each of them, we summarize the prominent methods and models, including approaches to mention encoding based on the self-attention architecture. Since many EL models take advantage of entity embeddings to improve their generalization capabilities, we provide an overview of the widely-used entity embedding techniques. We group the variety of EL approaches by several common research directions: joint entity recognition and linking, models for global EL, domain-independent techniques including zero-shot and distant supervision methods, and cross-lingual approaches. We also discuss the novel application of EL for enhancing word representation models like BERT. We systemize the critical design features of EL systems and provide their reported evaluation results.

1 Introduction

Entity Linking is the task of identifying an entity mention in *unstructured* text and establishing a link to an entry in a *structured* Knowledge Base (KB), such as Freebase (Bollacker et al., 2008), DBpedia (Auer et al., 2007), etc. It is an essential component of many information extraction and natural language understanding pipelines since it resolves the lexical ambiguity of named entities.

Neural networks have managed to excel in EL as in many other natural language processing tasks due to their ability to learn useful deep distributed representations of linguistic data (Collobert et al., 2011; Young et al., 2018). The state-of-the-art neural entity linking models have shown undoubted improvements over classical machine learning approaches based on feature engineering. In this survey, we systematize recently proposed neural models, distilling one generic architecture used by the majority of the neural EL models, but also discuss its prominent variations.

The important component of neural entity linking systems is entity vector representations and entity encoding methods. It has been shown that encoding the KB structure (entity relationships), entity definitions, as well as textual information in large unstructured corpora, helps to improve the generalization capabilities of EL models significantly. We summarize novel methods for entity encoding, as well as context/mention encoding techniques.

Many natural language processing systems take advantage of deep pre-trained language models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and their modifications. EL made its path into these models as a way of introducing information stored in KBs, which helps to adopt word representations to certain text processing tasks. We briefly discuss this novel application of EL.

There are few previous surveys devoted to the EL task (Rao et al., 2013; Ling et al., 2015; Shen et al., 2015; Al-Moslmi et al., 2020). In the most recent paper, Al-Moslmi et al. (2020) review both entity recognition and general entity disambiguation/linking methods published between the years 2014-2019. Instead, we focus specifically on rapidly developing neural models presented since 2015. The previous

* Equal contribution

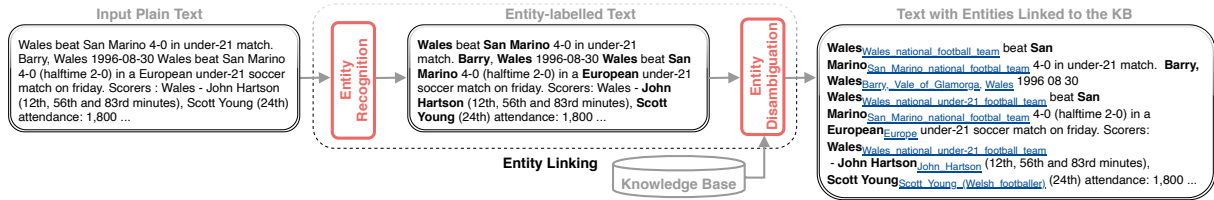


Figure 1: EL model takes a raw textual input and enriches it with entity mention links in a KB.

surveys do not address the topics of entity encoding, applications of EL to deep pre-trained language models, and cross-lingual EL. We also the first to summarize the domain-independent approaches to EL, many of which are based on zero-shot techniques. More specifically, this paper makes the following contributions:

- a survey of state-of-the-art neural entity linking models;
- a survey of entity embedding techniques;
- a discussion of recent domain-independent (zero-shot) and cross-lingual EL approaches;
- a survey of EL applications to modeling word representations.

The structure of this survey is the following. We start with defining the task of EL in Section 2. In Section 3.1, the common architecture of neural entity linking systems is presented. Modifications and variations of this basic pipeline are discussed in Section 3.2. In Section 4, we summarize the evaluation results for EL and entity representation models. Section 5 is dedicated to the recently emerged application of EL for improving neural language models. Finally, Section 6 summarizes the survey and suggests a prominent direction of future work in neural EL.

2 Task Description

Consider the example presented in Figure 1 with an entity mention *Scott Young*. Literally, this common name can at least refer to an *American football player*, *Welsh football player*, or a *writer*. The EL task is to correctly reveal the mention in the text, resolve the ambiguity, and provide a link to a corresponding entity entry in a KB. Wikification (Cheng and Roth, 2013) and Entity Disambiguation (ED) are considered as subtypes of EL (Navigli, 2009). In this survey, we assume that entity linking encompasses both entity recognition (ER) and entity disambiguation (ED). However, only few studies suggest models that perform ER and ED jointly, while the majority of papers referring EL focus only on ED and assume that mention boundaries are given by an external entity recogniser (Rizzo et al., 2014). ER techniques that perform only recognition without disambiguation are considered in many previous surveys (Nadeau and Sekine, 2007; Sharnagat, 2014; Goyal et al., 2018; Yadav and Bethard, 2018) and are out of the scope of this work.

To learn a mapping from entity mentions in a context to entity entries in a KB, EL models use supervision signals like manually annotated mention-entity pairs. The size of KBs vary; they can contain hundreds of thousands or millions of entities. Due to their large size, training data for EL would be extremely unbalanced; training sets can lack even a single example for a particular entity or mention, e.g. AIDA training set (Hoffart et al., 2011). To deal with this problem, EL models should have wide generalization capabilities. Despite their large size, KBs are incomplete. Therefore, some mentions in the text cannot be correctly mapped to any KB entry. Determining such unlikely mentions is one of the EL challenges. Furthermore, it is customary to distinguish “local” and “global” EL tasks. Local EL performs disambiguation of each mention in text independently using only context near target mentions, while global EL deals with simultaneous disambiguation of the all mentions and can engage features extracted from the whole document.

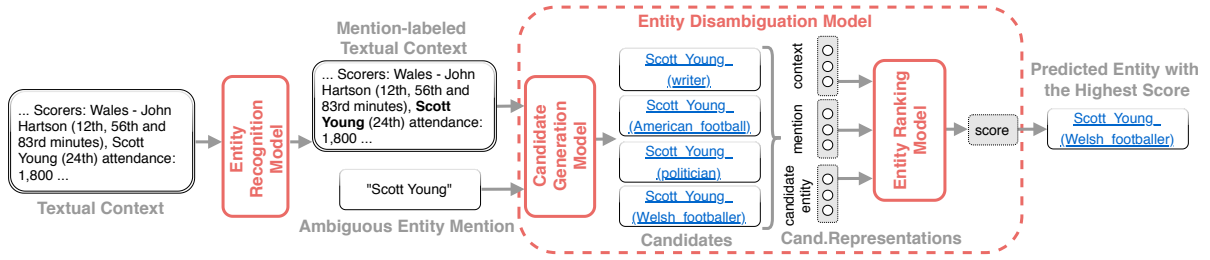


Figure 2: EL contains two main steps: *Entity Recognition*, mentions in a plain text are distinguished, and *Entity Disambiguation*, a corresponding entity is predicted for the given mention. Entity Disambiguation is further divided into two steps: *Candidate Generation*, possible entities are produced for the mention, and *Entity Ranking*, context/mention - candidate similarity score is computed through the representations.

3 Neural Entity Linking

We start the discussion of neural entity linking approaches from the most general structure of pipelines and continue with various specific modifications like joint entity recognition and linking, using global context, domain-independent approaches including zero-shot methods, and cross-lingual models. We also give a detailed overview of entity embedding techniques.

3.1 General Architecture

Some of the attempts to EL with neural networks treat it as a multi-class classification task, in which entities correspond to classes. However, the straightforward approach results in a large number of classes, which leads to suboptimal performance without task sharing (Kar et al., 2018). The streamlined approach to EL is to treat it as a ranking problem. We present the EL pipeline in Figure 2, which is applicable to the majority of the neural approaches. Here, the entity recognition model identifies the mention boundaries in text. The next step is to produce a short list of possible entities (candidates) for a mention. The entity ranking model estimates how well a candidate entity matches the context. An optional step is to determine unlinkable mentions, for which KBs do not contain a corresponding entity.

3.1.1 Candidate Generation

The goal of this step is given an ambiguous entity mention, such as “Scott Young”, to provide a list of its possible “senses” as specified by entities in a KB. EL is analogous to the Word Sense Disambiguation (WSD) task (Moro et al., 2014; Navigli, 2009) in terms of addressing lexical ambiguity. One of the major differences is that, in WSD, each sense of a word can be clearly defined by WordNet (Fellbaum, 1998), while, in EL, KBs do not provide such an exact mapping between mentions and entities. Therefore, a mention can be linked to any entity in a KB, resulting in large decision space, e.g. the notorious “Big Blue” for referring to IBM. To address this issue, preliminary filtering of an entity list, called candidate generation, is performed.

There are three prominent methods for this: a surface form matching, a dictionary lookup, and a prior probability computation. In the first approach, a candidate list is composed of entities, which simply match surface forms of mentions in the text (Zwickybauer et al., 2016; Moreno et al., 2017; Le and Titov, 2019b). For the example mention of “Big Blue”, this approach could not work well as the referent entity IBM does not contain a mention string. In the second approach, a dictionary of additional aliases is constructed using KB metadata like disambiguation/redirect pages of Wikipedia (Fang et al., 2019). Pershina et al. (2015) provide a resource of this type used in many EL models (Yamada et al., 2016; Cao et al., 2017; Newman-Griffis et al., 2018; Radhakrishnan et al., 2018; Martins et al., 2019). Another well-known alternative is the YAGO ontology (Suchanek et al., 2007) – automatically constructed from Wikipedia and WordNet. Among many other relations, it provides ‘*means*’ relations between mentions and entities, and this mapping is utilized as a candidate generator (Hoffart et al., 2011; Yamada et al., 2016; Ganea and Hofmann, 2017; Kolitsas et al., 2018; Cao et al., 2018; Peters et al., 2019; Yamada et al., 2020). In this technique, the external dictionaries would help to disambiguate “Big Blue” as IBM. In the

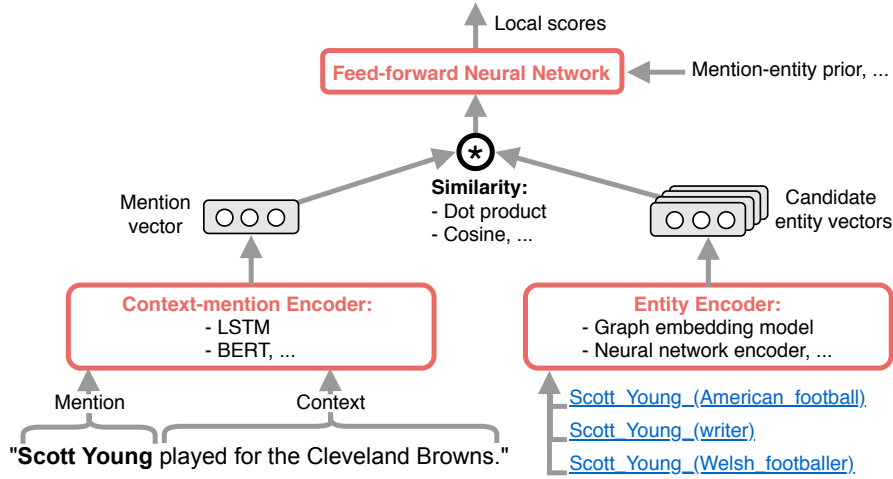


Figure 3: Generalized candidate entity ranking neural architecture.

third approach, the candidates are generated based on precalculated prior probabilities of correspondence between certain mentions and entities. These priors can be computed using the mention-entity hyperlink count statistics. Most of the studies rely on priors computed on the basis of Wikipedia anchor links (Zwiclbaauer et al., 2016; Tsai and Roth, 2016; Ganea and Hofmann, 2017; Kolitsas et al., 2018; Sil et al., 2018; Cao et al., 2018; Peters et al., 2019; Le and Titov, 2019a; Yang et al., 2019; Yamada et al., 2020). Another widely used option for this is CrossWikis (Spitkovsky and Chang, 2012), which is an extensive dictionary computed from the frequency of mention-entity links of web crawl data (Ganea and Hofmann, 2017; Gupta et al., 2017; Kolitsas et al., 2018; Cao et al., 2018; Peters et al., 2019; Yamada et al., 2020). The example mention string of “Big Blue” could be labeled as its referent entity IBM with precomputed priors. Recent zero-shot models (Logeswaran et al., 2019; Gillick et al., 2019; Wu et al., 2020) perform candidate generation without external knowledge. Section 3.2.3 describes them in detail.

3.1.2 Entity Ranking

The goal of this stage is given a list of entity candidates from a KB and a context with a mention to rank these entities assigning a score to each of them. Figure 3 depicts the typical architecture of the ranking component. To correctly disambiguate an entity mention, it is crucial to thoroughly capture the information from its context. A contextualized vector representation of a mention is generated by an encoder network. Several early techniques in neural EL utilize a convolutional encoder (Sun et al., 2015; Francis-Landau et al., 2016), as well as attention between candidate entity embeddings and embeddings of words surrounding a mention (Ganea and Hofmann, 2017; Le and Titov, 2019a). However, in recent models, two approaches prevail: recurrent networks and self-attention (Vaswani et al., 2017).

A recurrent architecture with LSTM cells (Hochreiter and Schmidhuber, 1997) that has been a backbone model for many NLP applications, is adopted to EL in (Martins et al., 2019; Kolitsas et al., 2018; Gupta et al., 2017; Sil et al., 2018; Le and Titov, 2019b; Fang et al., 2019). Gupta et al. (2017) concatenate outputs of two LSTM networks that independently encode left and right contexts of a mention (including the mention itself). In the same vein, Sil et al. (2018) encode left and right local contexts via LSTMs but also pool the results across all mentions in a coreference chain and postprocess left and right representations with a tensor network. A modification of LSTM – GRU (Chung et al., 2014) is used by Eshel et al. (2017) in conjunction with an attention mechanism (Bahdanau et al., 2015) to encode left and right context of a mention. Kolitsas et al. (2018) represent an entity mention as a combination of LSTM hidden states included in the mention span. Le and Titov (2019b) simply run a bidirectional LSTM network on words complemented with embeddings of word positions relative to a target mention. Shahbazi et al. (2019) adopt pre-trained ELMo (Peters et al., 2018) for mention encoding by averaging word representations inside mentions.

Encoding methods based on self-attention have recently become ubiquitous. The EL models presented in (Wu et al., 2020; Logeswaran et al., 2019; Peters et al., 2019; Yamada et al., 2020) rely on the outputs from pre-trained BERT layers (Devlin et al., 2019) for context and mention encoding. In Peters et al. (2019), a mention representation is modeled by pooling over word pieces in a mention span. The authors also put an additional self-attention block over all mention representations that encode interactions between several entities in a sentence. Another approach to modeling mentions is to insert special tags around them and perform a reduction of the whole encoded sequence. Wu et al. (2020) reduce a sequence by keeping the representation of the special pooling symbol ‘[CLS]’ inserted at the beginning of a sequence. Logeswaran et al. (2019) mark positions of a mention span by summing embeddings of words within the span with a special vector and use the same reduction strategy as Wu et al. (2020). Yamada et al. (2020) concatenate text with all mentions in it and jointly encode this sequence via a self-attention model based on pre-trained BERT.

The produced mention representation is compared with candidate entity representations. Entity representations can be pre-trained (see Section 3.1.3) or generated by another encoder as in some zero-shot approaches (see Section 3.2.3). The BERT-based model of Yamada et al. (2020) simultaneously learns how to encode mentions and entity embeddings in the unified architecture. Most of the state-of-the-art studies compare mention and entity representations using a dot product (Ganea and Hofmann, 2017; Gupta et al., 2017; Kolitsas et al., 2018; Peters et al., 2019; Wu et al., 2020) or cosine similarity (Sun et al., 2015; Francis-Landau et al., 2016; Gillick et al., 2019). The calculated similarity score is often combined with mention-entity priors obtained during the candidate generation phase (Francis-Landau et al., 2016; Ganea and Hofmann, 2017; Kolitsas et al., 2018) or other features including various similarities, string matching indicator, and entity types (Francis-Landau et al., 2016; Sil et al., 2018; Shahbazi et al., 2019; Yang et al., 2019). One of the common techniques for that is to use an additional one or two-layer feedforward network (Francis-Landau et al., 2016; Ganea and Hofmann, 2017; Shahbazi et al., 2019). The final disambiguation decision is inferred via a probability distribution, which is usually approximated by a softmax function over the candidates. The local similarity score or a probability distribution can be further utilized for global scoring (see Section 3.2.2).

The objective function is often formulated in terms of a ranking loss instead of the cross-entropy that is common for classification tasks. The idea behind such an approach is to enforce a positive margin between similarity scores of mentions to positive and negative candidates (Ganea and Hofmann, 2017; Kolitsas et al., 2018).

3.1.3 Entity Representations

The linking decision requires to measure how accurately candidate entities match a corresponding mention or context based on a structured or textual information of candidate entities. Low-dimensional semantic representations of entities encode this information in such a way that spatial proximity of entities in a vector space correlates with their semantic similarity. This fact is illustrated in Figure 4 for four entities of *Scott Young*. All four entities could be disambiguated with their closest entities; the most similar entities of *Scott Young (American football)* are related to American football including *Alex Henery*. In contrast, for *Scott Young (politician)*, the most similar entities are politicians.

The earliest methods, including Milne and Witten (2008), He et al. (2013), and Huang et al. (2015), depend on hand-engineered features such as bag-of-words or one-hot vectors to represent entities. After word2vec (Mikolov et al., 2013) gained success in word representations, its architecture was modified to be able to produce entity vectors. There are three common ways to apply this adjustment. The first one is to extend the objective function with a joint alignment function based on several features of entities, however, this requires sparse entity-entity co-occurrences statistics (Fang et al., 2016; Yamada et al., 2016; Cao et al., 2017; Shi et al., 2020). To overcome this issue, the second approach provides a formulation of the objective function based on an extensive entity-word co-occurrences statistics (Ganea and Hofmann, 2017; Radhakrishnan et al., 2018). The last approach is to directly replace the raw input text with entity annotated text without any statistics (Zwicklbauer et al., 2016; Moreno et al., 2017; Tsai and Roth, 2016).

More recently, Gupta et al. (2017) aim at capturing different kinds of entity information, including

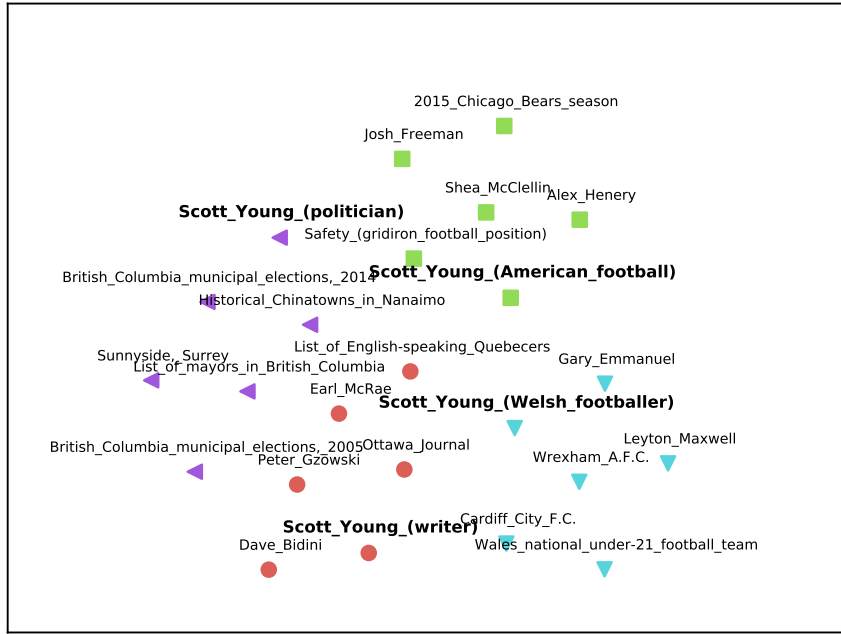


Figure 4: Entity embedding space for entities related to the ambiguous entity mention “Scott Young”. Four candidate entities from Wikipedia/DBpedia are illustrated. For each entity, their most similar 5 entities are shown in the same colors. Entity embeddings are visualized with t-SNE using pre-trained embeddings provided by Sevgili et al. (2019).

entity type, description page, linked mention, and contextual information, and therefore, they generate a large encoder, which involves CNN for the entity description and alignment function for the others. Gillick et al. (2019) encode entities based on their title, description page, and category information. All previously mentioned models rely on the annotated data, and a few studies are challenged with less resource dependence. In this track, Sun et al. (2015) and Sil et al. (2018) derive entity embeddings using pre-trained word2vec word vectors through description page words, surface forms words, and entity category words (Sun et al. (2015) continue learning representations while disambiguation and they use annotation in the disambiguation phase). Newman-Griffis et al. (2018) expand the word2vec architecture with a distant supervision setup based on the terminology of Wikipedia’s page titles and redirects. Sevgili et al. (2019) build a graph from entity-entity hyperlinks and use a graph embedding method to generate entity embeddings. Logeswaran et al. (2019) and Wu et al. (2020) depend on the BERT architecture to create representations through the description pages.

The neural architectures for learning representations are mostly designed to keep word and entity in joint semantic space to allow a straightforward ranking of candidates with vector similarity. They are trained with three common types of score functions. The first and most commonly used one is the similarity or the prior score of entity and mention. The second direction is the alignment score function, where several features are learned independently and joined in the alignment function (Yamada et al., 2016; Fang et al., 2016; Cao et al., 2017; Gupta et al., 2017). The final one relies on cross- or bi-encoders of entity and mention/context, which are built over the BERT architecture (Logeswaran et al., 2019; Wu et al., 2020). The representation models also vary in terms of data sources, which can be structured (e.g. hyperlinks) or textual (i.e. description of an entity, anchor texts, or annotated texts). The detailed model-wise comparison can be found in Table 3 in the appendix.

3.1.4 Unlinkable Mention Prediction

The referent entities of some mentions can be absent in the KBs, e.g. there is no Wikipedia entry about *Scott Young* as a cricket player of the Stenhousemuir cricket club.¹ Therefore, an EL system should be able to predict the absence of a reference, which is known as NIL prediction. There are four common ways to perform NIL prediction. Sometimes a candidate generator does not yield any corresponding entities for a mention; such mentions are trivially considered unlinkable (Tsai and Roth, 2016; Sil et al., 2018). One can set a threshold for the best linking probability (or a score), below which mention is considered unlinkable (Peters et al., 2019; Lazic et al., 2015). Some models introduce an additional special ‘NIL’ entity in the ranking phase, so models can predict it as the best match for the mention (Kolitsas et al., 2018). It is also possible to train an additional binary classifier that accepts mention-entity pairs after the ranking phase, as well as several additional features (best linking score, whether mentions are also detected by a dedicated NER system, etc.), and makes the final decision about whether a mention is linkable or not (Moreno et al., 2017; Martins et al., 2019).

3.2 Modifications of the General Architecture

This section presents the most notable modifications and improvements of the general architecture of neural entity linking models presented in Section 3.1 and Figures 2 and 3.

3.2.1 Joint Entity Recognition and Disambiguation Architectures

A few systems provide a joint solution for entity recognition and linking. Undoubtedly, solving these two problems simultaneously makes the task more challenging. However, the interaction between these steps can be beneficial for improving the quality of the overall pipeline due to their natural mutual dependency. While first competitive models that provide joint solution were probabilistic graphical models (Luo et al., 2015; Nguyen et al., 2016), we focus on purely neural approaches proposed recently (Kolitsas et al., 2018; Peters et al., 2019; Martins et al., 2019; Broscheit, 2019).

The main difference of joint pipelines is the necessity to produce also mention candidates. For this purpose, (Peters et al., 2019; Kolitsas et al., 2018) enumerate all spans in a sentence with a certain maximum width, filter them by several heuristics (remove mentions with stop words, punctuation, ellipses, quotes, and currencies), and try to match them to a pre-built index of entities used for the candidate generation. If a mention candidate has at least one corresponding entity candidate, it is further treated by a ranking neural network that can also discard it by considering it unlinkable to any entity in a KB (see Section 3.1.2). Therefore, the decision during the entity disambiguation phase affects entity recognition.

Martins et al. (2019) describe the approach with tighter integration between recognition and linking phases via multi-task learning. The authors propose a stack-based bidirectional LSTM network with a shift-reduce mechanism and attention for entity recognition that propagates its internal states to the linker network for candidate entity ranking. The linker is supplemented with a NIL predictor network. The networks are trained jointly by optimizing the sum of losses from all three components.

Broscheit (2019) goes further by suggesting a completely end-to-end method that deals with entity recognition and linking jointly without explicitly executing a candidate generation step. They formulate the task as a sequence labeling problem, where each token in a text is assigned an entity link or a NIL class. They leverage a sequence tagger based on pre-trained BERT for this purpose. This simplistic approach does not supersede (Kolitsas et al., 2018) but outperforms the baseline, in which candidate generation, entity recognition, and linking are performed independently.

3.2.2 Global Context Architectures

Two kinds of context information are accepted to perform entity disambiguation: local and global. Local approaches rely on the words around the entity mention in a specified window, and each mention is disambiguated independently. This kind of method does not perform properly if the surrounding words do not carry sufficient contextual information (Fang et al., 2019). Semantic consistency across all entities in a context is also quite informative for the disambiguation. Global approaches use this topical coherence

¹<https://www.stenhousemuircricketclub.com/teams/171906/player/scott-young-1828009>

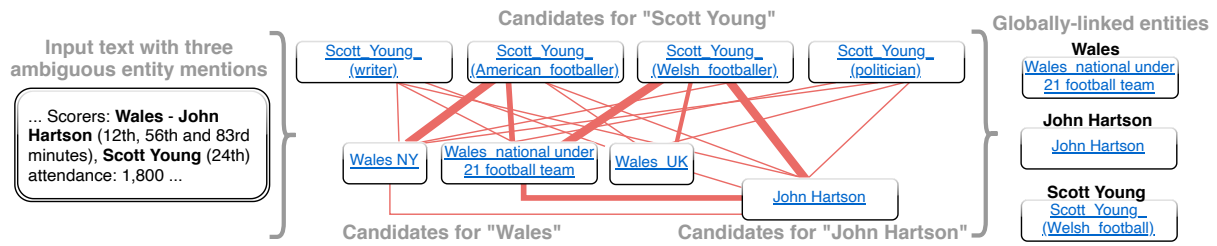


Figure 5: The global entity linking resolves all mentions simultaneously based on the coherence of entities. Bolder lines indicate higher degrees of entity-entity similarity.

and disambiguate all mentions simultaneously, as illustrated in Figure 5. Although the extra information of the global context improves the disambiguation accuracy, the number of possible entity assignments is combinatorial, which results in a high time complexity (Yang et al., 2019). Another difficulty is to attempt to assign an entity with its coherence score, since this score is not possible to compute in advance due to the simultaneous disambiguation (Yamada et al., 2016).

Earlier, global EL approaches typically perform PageRank or RandomWalk algorithms over a graph containing candidate entities of mentions in a context (Zwickybauer et al., 2016; Pershina et al., 2015; Guo and Barbosa, 2018). Another well-known solution is to maximize the Conditional Random Fields score function, which contains two terms: one evaluates an entity-context propriety, and the other one measures coherence (Ganea and Hofmann, 2017; Le and Titov, 2018; Le and Titov, 2019a). However, the exact inference is NP-hard. Hofmann et al. (2017) adapt loopy belief propagation with message passing iterations using pairwise entity scores to reduce the complexity, and Le and Titov (2018) expand it with a coreference relation of mentions as latent variables (the mentions are coreferent if they refer to the same entity). Some recent studies define the global EL problem as a sequential decision task, where the disambiguation of new entities is based on the already disambiguated ones. Fang et al. (2019) apply LSTM to be able to maintain long term memory for previous decisions. Yang et al. (2019) execute Dynamic Context Augmentation, where basically previous decisions are collected as dynamic context to improve the following predictions, and Yamada et al. (2020) compute confidence scores based on previous predictions. Another alternative for the recent direction is to attach an entity relatedness score to the score function of the entire model. Kolitsas et al. (2018) first select a set of entities with a high local score and compute the similarity between the in-process entity embedding and an average of the selected entity embeddings. Fang et al. (2016) calculate the similarity between the present entity and its surrounding entity candidates in a specified window. Yamada et al. (2016) and Radhakrishnan et al. (2018) measure the similarity first based on unambiguous mentions and then predict entities. Rather than computing entity score, Tsai and Roth (2016) directly append previous entity embeddings to the model. The final common approach is to rely on global information to capture the coherence instead of directly including an entity coherence component (Peters et al., 2019; Gupta et al., 2017; Cao et al., 2017; Moreno et al., 2017; Sil et al., 2018). Distinctively, Cao et al. (2018) integrate Graph Convolutional Network into a neural model to handle the global information, which is represented as a subgraph of candidate entities.

3.2.3 Domain-Independent Architectures

Domain independence is one of the most desired properties of EL systems. Annotated resources are very limited and exist only for a few domains. Obtaining labeled data in a new domain requires much labor. Earlier, this problem is tackled by few domain-independent approaches based on unsupervised (Wang et al., 2015; Cao et al., 2017; Newman-Griffis et al., 2018; Le and Titov, 2018) and semi-supervised models (Lazic et al., 2015). Recent studies provide solutions based on distant learning and zero-shot methods.

The studies (Le and Titov, 2019b; Le and Titov, 2019a) propose distant learning techniques that use only unlabeled documents. They rely on the weak supervision coming from a surface matching heuristic, and the EL task is framed as binary multi-instance learning. The algorithm learns to distinguish between

positive entities set and random negatives set. The positive set is obtained by retrieving entities with high word overlap with entities and relations to other mention candidates in the sentence. While showing promising performance, which in some cases rivals fully supervised systems, these approaches require either a KB describing relations of entities (Le and Titov, 2019b) or mention-entity priors computed from entity hyperlink statistics extracted from Wikipedia (Le and Titov, 2019a).

Recently, zero-shot techniques have seen a surge in EL (Logeswaran et al., 2019; Wu et al., 2020; Gillick et al., 2019; Gupta et al., 2017). The zero-shot setting means that only descriptions of entities are available, while other sources of supervision such as relations between entities in a Knowledge Graph (KG) and entity types are absent. This allows applying the system to almost any source of entities, since building a database of entities and their descriptions might be far less laborious compared to the development of complex KGs with the abundance of relations.

Since pre-build resources for candidate generation are not available for the target domain, for candidate selection, one can rely only on textual descriptions of entities. Logeswaran et al. (2019) use the BM25 information retrieval formula (Jones et al., 2000) to rank entity Wikipedia pages via mentions and keep entities with top relevant pages as candidates. Wu et al. (2020) use the BERT bi-encoder on top of entity descriptions to select candidates. The bi-encoder consists of two networks that separately encode context and an entity description. As in supervised approaches, entity and mention representations are compared via a dot product, and the top candidates are selected for ranking. In Gillick et al. (2019), the candidate generation stage is absent at all, and ranking is performed for all entities in a KG.

For ranking entities, which are absent in a training corpus of the source domain, zero-shot methods rely on entity description encoding robust across multiple domains. One of the first studies that utilize such a technique is proposed by Gupta et al. (2017) (not purely zero-shot because they used entity typings). Gillick et al. (2019) propose a CNN network for encoding entity descriptions along with optional entity typing information. It is also worth noting that Gillick et al. (2019) rely on annotated data while training entity representations. Some other approaches (Logeswaran et al., 2019; Wu et al., 2020) utilize the BERT-based cross-encoder to perform joint encoding of mentions and entities. The cross-encoder takes concatenation of context with a mention and an entity description to produce a scalar score for each candidate. In both studies, cross-encoders achieve substantially better results compared to other approaches.

3.2.4 Cross-lingual Architectures

The cross-lingual EL methods (Ji et al., 2015) leverage supervision signals from multiple languages for training a model in a target language. For example, the inter-lingual links in Wikipedia can be utilized for alignment of entities in multiple languages. Using such an alignment, the annotated data from high-resource languages like English can help to improve the quality of text processing for the low-resource ones.

One of the challenges in cross-lingual EL is candidate generation, since the low-resource language can lack mappings between mention strings and entities. In this case, candidate generation can be approached by: mining a translation dictionary (Pan et al., 2017), training a translation and alignment model (Tsai and Roth, 2018), or applying a neural character level string matching model (Rijhwani et al., 2019). The latter relies on training on a high-resource pivot language, similar to the target low-resource one. The neural string matching approach can be further improved with simpler average n-gram encoding and extending entity-entity pairs with mention-entity examples (Zhou et al., 2020).

There are several approaches to candidate ranking that take advantage of cross-lingual data for dealing with the lack of annotated examples. Pan et al. (2017) uses the comparison of Abstract Meaning Representation (AMR) (Banarescu et al., 2013) statistics in English Wikipedia and mention context for ranking. To train an AMR tagger, pseudo-labeling (Lee, 2013) was used. Tsai and Roth (2016) train monolingual embeddings for words and entities jointly by replacing every entity mention with corresponding entity tokens. Using the connection of entities to Wikipedia pages that exist for multiple languages, they learn the projection functions from multiple languages into the English embedding space. For ranking, context embeddings are averaged, projected into English space, and compared with entity embeddings. The authors demonstrate that this approach helps to build better entity representations and boost EL per-

formance on low-resource languages. Sil et al. (2018) propose a method for zero-shot transfer from a high-resource language. The authors extend the previous approach with the least squares objective for embedding projection learning, the CNN context encoder, and a trainable re-weighting of each dimension of context and entity representations. The proposed approach demonstrates improved performance compared to previous non-zero-shot approaches. Upadhyay et al. (2018) argued that the success of zero-shot cross-lingual approaches might be highly related to mention-entity prior probabilities used as features. Their approach extends (Sil et al., 2018) with global context information and incorporation of typing information into context and entity representations (the system learns to predict typings during the training). The authors report a significant drop in performance for zero-shot cross-lingual EL with an excluded mention-entity prior. They also show that training on the high-resource language might be very beneficial for the low-resource settings.

3.3 Summary

We summarize critical design features for neural EL models in Table 4 in the appendix. One can note, the EL systems do not utilize the whole spectrum of available features. The mention encoders have made a shift to self-attention architectures and start using deep pre-trained models like BERT. The majority of studies still rely on external knowledge for the candidate generation step. There is a surge of models that tackle the domain adaptation problem in a zero-shot fashion. However, the task of zero-shot joint entity recognition and linking has not been addressed yet. It is shown in several works that the cross-encoder architecture is superior compared to models with separate mention and entity encoders. Many approaches rely on pre-trained entity representations, only few take advantage of a trainable entity encoder inside an EL model. The global context is widely used, but there are few recent studies that focus only on local EL.

4 Evaluation

In this section, we present and summarize the evaluation results for entity embeddings and neural entity linking systems reported by their authors.

4.1 Entity Relatedness

An entity relatedness dataset was put forward by Ceccarelli et al. (2013) using the dataset of Hoffart et al. (2011). The dataset is in the form of queries, where the first entity is accepted as correctly linked and the second entity is the candidate. Here, the evaluation task is to rank entities for the target one, which is performed using cosine similarity of entity representations except for two studies: Milne and Witten introduce a Wikipedia hyperlink-based measure, known as WLM, and recently El Vaigh et al. (2019) provide a weighted semantic relatedness measure. The evaluation of ranking quality is performed with normalized discounted cumulative gain (nDCG) (Järvelin and Kekäläinen, 2002) and mean average precision (MAP) (Yue et al., 2007). nDCG is a well-known quality metric used in information retrieval, which provides a fair evaluation by measuring the position impressiveness. Similarly, MAP measures how accurately the model performs for the target entity.

The highest score is reported by Huang et al. (2015); they specifically train the embeddings based on a pairwise entity score function in a supervised way. Instead, in other models, entity embeddings are trained in a joint space with word embeddings based on the relatedness between mentions and words. Therefore, the results of Huang et al. (2015) are distinctively higher. Ganea and Hofmann (2017) and Cao et al. (2017) achieve good scores, and recently, Shi et al. (2020) also present an excellent performance using a large amount of data sources based on textual and KB information.

4.2 Entity Linking

We report EL performance results on widely-used datasets: AIDA (Hoffart et al., 2011), TAC KBP 2010 (Ji et al., 2010), MSNBC (Cucerzan, 2007), AQUAINT (Milne and Witten, 2008), ACE2004 (Ratinov et al., 2011), CWEB (Guo and Barbosa, 2018; Gabrilovich et al., 2013), and WW (Guo and Barbosa, 2018; Gabrilovich et al., 2013), in Table 2. Among them, CWEB and WW are the largest datasets that

	nDCG@1	nDCG@5	nDCG@10	MAP
Milne and Witten (2008)	0.540	0.520	0.550	0.480
Huang et al. (2015)	0.810	0.730	0.740	0.680
Yamada et al. (2016)	0.590	0.560	0.590	0.520
Ganea and Hofmann (2017)	0.632	0.609	0.641	0.578
Cao et al. (2017)	0.613	0.613	0.654	0.582
El Vaigh et al. (2019)	0.690	0.640	0.580	-
Shi et al. (2020)	0.680	0.814	0.820	-

Table 1: Reported results for entity relatedness evaluation on the dataset of Ceccarelli et al. (2013) .

are annotated automatically, while AIDA is the largest dataset, annotated manually. AIDA contains the development set AIDA-A and the test set AIDA-B. We report the results calculated for AIDA-B. Some of these results are evaluated using GERBIL (Usbeck et al., 2015) – a benchmarking platform for entity recognition and disambiguation systems. The cross-lingual EL results are reported for the TAC KBP 2015 (Ji et al., 2015) Spanish (es) and Chinese (zh) datasets. We present accuracy and micro F1 scores. The micro F1 scores for systems that perform ER and EL jointly are different from the measures reported for disambiguation only systems due to mistakes in entity recognition.

	AIDA-B		KBP'10	MSNBC	AQUAINT	ACE-2004	CWEB	WW	KBP'15 (es)	KBP'15 (zh)
	Accuracy	Micro F1	Accuracy	Micro F1	Micro F1	Micro F1	Micro F1	Micro F1	Accuracy	Accuracy
Sun et al. (2015)	-	-	0.839	-	-	-	-	-	-	-
Lazic et al. (2015)	0.864	-	-	-	-	-	-	-	-	-
Tsai and Roth (2016)	-	-	-	-	-	-	-	-	0.809	0.836
Fang et al. (2016)	-	-	0.889	0.755	0.852	0.808	-	-	-	-
Yamada et al. (2016)	0.915	-	0.855	-	-	-	-	-	-	-
Zwicklbauer et al. (2016)	0.784	-	-	0.911	0.842	0.907	-	-	-	-
Francis-Landau et al. (2016)	0.855	-	-	-	0.899	-	-	-	-	-
Eshel et al. (2017)	0.833	-	-	-	-	-	-	-	-	-
Ganea and Hofmann (2017)	0.922	-	-	0.937	0.885	0.885	0.779	0.775	-	-
Gupta et al. (2017)	0.829	-	-	-	-	0.907	-	-	-	-
Cao et al. (2017)	0.890	-	-	-	-	-	-	-	-	-
Newman-Griffis et al. (2018)	0.639	-	-	-	-	-	-	-	-	-
Sil et al. (2018)	0.940	-	0.874	-	-	-	-	-	0.823	0.844
Kolitsas et al. (2018)	-	0.824	-	0.724	-	-	-	-	-	-
Le and Titov (2018)	0.931	-	-	0.939	0.883	0.899	0.775	0.780	-	-
Radhakrishnan et al. (2018)	-	-	0.896	-	-	-	-	-	-	-
Cao et al. (2018)	0.800	-	0.910	-	0.870	0.880	-	0.860	-	-
Raiman and Raiman (2018)	0.949	-	0.909	-	-	-	-	-	-	-
Upadhyay et al. (2018)	-	-	-	-	-	-	-	-	0.535	0.559
Gillick et al. (2019)	-	-	0.870	-	-	-	-	-	-	-
Le and Titov (2019b)	0.815	-	-	-	-	-	-	-	-	-
Martins et al. (2019)	-	0.819	-	-	-	-	-	-	-	-
Peters et al. (2019)	-	0.744	-	-	-	-	-	-	-	-
Le and Titov (2019a)	0.897	-	-	0.922	0.907	0.881	0.782	0.817	-	-
Fang et al. (2019)	0.943	-	-	0.928	0.875	0.912	0.785	0.828	-	-
Yang et al. (2019)	0.946	-	-	0.946	0.874	0.894	0.735	0.782	-	-
Shahbazi et al. (2019)	0.935	-	0.883	-	-	-	-	-	-	-
Broscheit (2019)	-	0.793	-	-	-	-	-	-	-	-
Wu et al. (2020)	-	-	0.940	-	-	-	-	-	-	-
Yamada et al. (2020)	0.950	-	-	0.963	0.935	0.919	0.789	0.891	-	-

Table 2: Reported results for entity disambiguation/linking evaluation on various datasets.

Among the joint recognition and disambiguation solutions, the leadership is still owned by Kolitsas et al. (2018). This system and others that solve the ER task fall behind the disambiguation-only systems since they rely on noisy mention boundaries produced by themselves. Among published local models for disambiguation, the best result is reported by Wu et al. (2020). It is worth noting that this model can be used in a zero-shot setting. The global models expectedly outperform the local ones. The work of Yamada et al. (2020) reports results that are consistently better compared to other solutions. The performance improvements are attributed to the masked entity prediction mechanism for entity embedding and the use of the pre-trained model based on BERT with the interdependent scoring function.

5 Applications of Entity Linking for Training Word Representation Models

Entity linking is an important component for solving such text processing tasks as semantic parsing (Berant and Liang, 2014), information extraction (Hoffmann et al., 2011), and question answering (Yih et al., 2015). In addition to such end tasks, a new trend is the use EL information for representation learning. Namely, several studies have shown that contextual word representations could benefit from

information stored in KBs by incorporating EL into deep models for transfer learning.

KnowBERT (Peters et al., 2019) injects between top layers of the BERT architecture one or several entity linkers and optimizes the whole network for multiple tasks: the masked language model (MLM) task, next sentence prediction, and EL. The authors adopt the general end-to-end EL pipeline of (Kolitsas et al., 2018) but use only the local context for disambiguation and use an encoder based on self-attention over the representations generated by underlying BERT layers. If the EL subsystem detects an entity mention in a given sentence, corresponding pre-built entity representations of candidates are utilized for calculating the updated contextual word representations generated on the current BERT layer. These representations are used as input in a subsequent layer and can also be modified by a subsequent EL subsystem. Experiments with two EL subsystems based on Wikidata and WordNet show that presented modifications in KnowBERT help it to slightly surpass other deep pre-trained language models in tasks of relationship extraction, WSD, and entity typing.

ERNIE (Zhang et al., 2019) expands the BERT (Devlin et al., 2019) architecture with a knowledgeable encoder (K-Encoder), which fuses contextualized word representations obtained from the underlying self-attention network with entity representations from a pre-trained TransE model (Bordes et al., 2013). EL in this study is performed by an external tool TAGME (Ferragina and Scaiella, 2010). For model pre-training, in addition to the MLM task, the authors introduce the task of restoring randomly masked entities in a given sequence keeping the rest of the entities and tokens. Using English Wikipedia and Wikidata as training data, the authors show that introduced modifications provide performance gains in entity typing, relation classification, and several GLUE tasks.

Wang et al. (2019) propose to train a disambiguation network using the composition of two losses: regular MLM and a Knowledge Embedding (KE) loss based on TransE (Bordes et al., 2013) objective for encoding graph structures. In KE loss, representations of entities are obtained from their textual descriptions encoded with a self-attention network (Liu et al., 2019), and representations of relations are trainable vectors. Although the system exhibits a significant drop in performance on general NLP benchmarks such as GLUE (Wang et al., 2018), it shows increased performance on a wide range of KB-related tasks such as TACRED (Zhang et al., 2017), FewRel (Han et al., 2018), and OpenEntity (Choi et al., 2018).

6 Conclusion and Future Directions

In this survey, we have analyzed the recently proposed neural entity linking models. The majority of studies still rely on external knowledge for the candidate generation step. The mention encoders have made a shift to self-attention architectures and start using deep pre-trained models like BERT. There is a surge of models that tackle the domain adaptation problem in a zero-shot fashion. It is shown in several works that the cross-encoder architecture is superior compared to models with separate mention and entity encoders. Many approaches rely on pre-trained entity representations, only few take advantage of a trainable entity encoder inside an EL model. The global context is widely used, but there are few recent studies that focus only on local EL. Among the joint recognition and disambiguation solutions, the leadership is still owned by Kolitsas et al. (2018). Among published local models for disambiguation, the best result is reported by Wu et al. (2020). It is worth noting that this model can be used in a zero-shot setting. The global models expectedly outperform the local ones. The work of Yamada et al. (2020) reports results that are consistently better compared to other solutions. The performance improvements are attributed to the masked entity prediction mechanism for entity embedding and to the usage of the pre-trained model based on BERT with the interdependent scoring function. It is also worth noting that several studies have demonstrated some benefits for deep transfer learning models of using information stored in KBs by incorporating EL into these models.

In the future work, we expect that zero-shot EL will rapidly evolve engaging other features like global coherence across all entities in a document, NIL prediction, joining ER and EL steps together, or providing completely end-to-end solutions. The latter would be an especially challenging task but also a fascinating research direction.

References

- Tareq Al-Moslmi, Marc Gallofr Ocaa, Andreas L. Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8:32862–32881.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC07/ASWC07, page 722735, Berlin, Heidelberg. Springer-Verlag.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, San-Diego, California, USA.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland. Association for Computational Linguistics.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD 08, page 12471250, New York, NY, USA. Association for Computing Machinery.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, Stateline, Nevada, USA.
- Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.
- Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1623–1633, Vancouver, Canada. Association for Computational Linguistics.
- Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Neural collective entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 675–686, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2013. Learning relatedness measures for entity linking. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, CIKM '13*, pages 139–148, New York, NY, USA. ACM.
- Xiao Cheng and Dan Roth. 2013. Relational inference for Wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Seattle, Washington, USA. Association for Computational Linguistics.
- Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.
- Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*, Montral, Canada.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.
- Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Cheikh Brahim El Vaigh, François Goasdoué, Guillaume Gravier, and Pascale Sébillot. 2019. Using knowledge base semantics in context-aware entity linking. In *Proceedings of the ACM Symposium on Document Engineering 2019, DocEng 19*, New York, NY, USA. ACM.
- Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.
- Wei Fang, Jianwen Zhang, Dilin Wang, Zheng Chen, and Ming Li. 2016. Entity disambiguation by knowledge and text jointly embedding. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 260–269, Berlin, Germany. Association for Computational Linguistics.
- Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The World Wide Web Conference, WWW '19*, pages 438–447, New York, NY, USA. ACM.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.
- Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-Fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 10*, page 16251628, New York, NY, USA. ACM.
- Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261, San Diego, California, USA.
- Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. FACC1: Freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). Note: <http://lemurproject.org/clueweb09/>.
- Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.
- Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldrige, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.
- Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.
- Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.
- Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.
- Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.
- Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–34, Sofia, Bulgaria. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenu, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 782–792. Association for Computational Linguistics.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.
- Hongzhao Huang, Larry P. Heck, and Heng Ji. 2015. Leveraging deep neural networks and knowledge graphs for entity disambiguation. *CoRR*, abs/1504.07678.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422446, October.
- Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Third Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.
- Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015*, pages 16–17, Gaithersburg, Maryland, USA. NIST.
- Karen Sprck Jones, Shelia Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments part 2. *Information Processing & Management*, 36(6):809840.
- Rijula Kar, Susmija Reddy, Sourangshu Bhattacharya, Anirban Dasgupta, and Soumen Chakrabarti. 2018. Task-specific representation learning for web-scale entity disambiguation. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5812–5819, New Orleans, Louisiana, USA. AAAI Press.
- Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.
- Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. 2015. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515.
- Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.
- Phong Le and Ivan Titov. 2019a. Boosting entity linking performance by leveraging unlabeled documents. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1935–1945, Florence, Italy. Association for Computational Linguistics.
- Phong Le and Ivan Titov. 2019b. Distant learning for entity linking with automatic noise detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4081–4090, Florence, Italy, July. Association for Computational Linguistics.
- Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, Atlanta, USA. JMLR.
- Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.
- Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal.

- Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS13, page 31113119, Red Hook, NY, USA. Curran Associates Inc.
- David Milne and Ian H. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.
- Jose G. Moreno, Romaric Besançon, Romain Beaumont, Eva D'hondt, Anne-Laure Ligozat, Sophie Rosset, Xavier Tannier, and Brigitte Grau. 2017. Combining word and entity embeddings for entity linking. In *Extended Semantic Web Conference (1)*, volume 10249 of *Lecture Notes in Computer Science*, pages 337–352.
- Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.
- Denis Newman-Griffis, Albert M. Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206, Melbourne, Australia. Association for Computational Linguistics.
- Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2016. J-NERD: joint named entity recognition and disambiguation with rich linguistic features. *Transactions of the Association for Computational Linguistics*, 4:215–229.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized page rank for named entity disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–243, Denver, Colorado, USA. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2227–2237, New Orleans, Louisiana, USA. AAAI Press.
- Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.
- Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. 2018. ELDEN: Improved entity linking using densified knowledge graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1844–1853, New Orleans, Louisiana. Association for Computational Linguistics.
- Jonathan Raiman and Olivier Raiman. 2018. Deeptype: Multilingual entity linking by neural type system evolution. In *AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA., Feb.
- Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

- Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931, Honolulu, Hawaii, USA, January.
- Giuseppe Rizzo, Marieke van Erp, and Raphaël Troncy. 2014. Benchmarking the extraction and disambiguation of named entities on the semantic web. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4593–4600, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
- Özge Sevgili, Alexander Panchenko, and Chris Biemann. 2019. Improving neural entity disambiguation with graph embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 315–322, Florence, Italy. Association for Computational Linguistics.
- Hamed Shahbazi, Xiaoli Z Fern, Reza Ghaeini, Rasha Obeidat, and Prasad Tadepalli. 2019. Entity-aware elmo: Learning contextual entity representation for entity disambiguation. *arXiv preprint arXiv:1908.05762*.
- Rahul Sharnagat. 2014. Named entity recognition: A literature survey. *Center For Indian Language Technology*.
- Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *Transactions on Knowledge & Data Engineering*, 27(2):443–460.
- Wei Shi, Siyuan Zhang, Zhiwei Zhang, Hong Cheng, and Jeffrey Xu Yu. 2020. Joint embedding in named entity linking on sentence level. *ArXiv*, abs/2002.04936.
- Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA. AAAI Press.
- Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3168–3175, Istanbul, Turkey. European Language Resources Association (ELRA).
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 07*, page 697706, New York, NY, USA. ACM.
- Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling mention, context and entity with neural networks for entity disambiguation. In *Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15*, pages 1333–1339. AAAI Press.
- Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual Wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California, USA. Association for Computational Linguistics.
- Chen-Tse Tsai and Dan Roth. 2018. Learning better name translation for cross-lingual Wikification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA. AAAI Press.
- Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. 2015. GERBIL: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pages 1133–1143, Florence, Italy. International World Wide Web Conferences Steering Committee.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 60006010, Red Hook, NY, USA. Curran Associates Inc.
- Han Wang, Jin Guang Zheng, Xiaogang Ma, Peter Fox, and Heng Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 695–704, Lisbon, Portugal. Association for Computational Linguistics.

- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. Kepler: A unified model for knowledge embedding and pre-trained language representation. *arXiv preprint arXiv:1911.06136*.
- Ledell Yu Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Zero-shot entity linking with dense entity retrieval. *ArXiv*, abs/1911.03814.
- Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, NM, USA, August. Association for Computational Linguistics.
- Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.
- Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2020. Global entity disambiguation with pretrained contextualized embeddings of words and entities. *arXiv preprint arXiv:1909.00426v2*.
- Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. Learning dynamic context augmentation for global entity linking. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 271–281, Hong Kong, China. Association for Computational Linguistics.
- Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.
- Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75.
- Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 07, page 271278, New York, NY, USA. Association for Computing Machinery.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, Copenhagen, Denmark. ACL.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.
- Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics*, 8:109–124.
- Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016. Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’16, pages 425–434, New York, NY, USA. ACM.

A Entity Embeddings

	annotated text	entity-entity links	entity-mention links	entity descriptions	entity titles	entity types/redirects	initialized	joint space entity and mention
Huang et al. (2015)		✓	✓	✓		✓	BoW one-hot	only entity
Sun et al. (2015)	✓				✓	✓	word2vec	✓ ¹
Fang et al. (2016)	✓	✓	✓	✓			-	✓
Yamada et al. (2016)	✓	✓					-	✓
Zwacklbauer et al. (2016)	✓ ²			✓			-	only entity
Tsai and Roth (2016)	✓				✓		-	✓
Ganea and Hofmann (2017)	✓						word2vec	✓
Cao et al. (2017)	✓	✓	✓				-	✓
Moreno et al. (2017)	✓						-	✓
Gupta et al. (2017)	✓			✓		✓	GloVe	✓ ⁴
Sil et al. (2018)				✓			word2vec	✓
Upadhyay et al. (2018)	✓		✓			✓	-	✓
Newman-Griffis et al. (2018)					✓	✓	-	✓
Radhakrishnan et al. (2018)	✓							✓
Rijhwani et al. (2019)	✓	✓			✓		-	✓
Logeswaran et al. (2019)				✓			BERT	✓ ³
Gillick et al. (2019)	✓			✓	✓	✓	GLove	✓
Sevgili et al. (2019)		✓		✓			-	only entity
Shi et al. (2020)	✓	✓				✓	-	✓
Zhou et al. (2020)	✓	✓	✓		✓		-	✓
Wu et al. (2020)				✓	✓		BERT	✓ ⁵

Table 3: Entity embedding models in terms of their data requirements and architectural features: the first six columns denote data related features; the remaining ones refer to the architectural features.

In Table 3, we present the EL models that generate entity representations for a part of their solutions. Annotated text data contain a sequence of terms, where entities are labeled. It is a powerful resource to catch textual information of an entity and a mention, however, many specialized domains are lacking such annotations (Newman-Griffis et al., 2018). Alternatively, it could be used as a sequence of entities by removing all words in the context, as Zwacklbauer et al. (2016) did. Entity description pages are another text resource provided by KBs. They contain a textual description for each entity, however, it is unable to catch any relation, like mention-entity or entity-entity. Entity-entity links and entity-mention links are quite informative about the relational knowledge. Entity titles, types, and redirect pages are dictionary-like information, where redirect pages are the most useful feature in terms of entity-entity relations.

The following notations were used to present the external features of the model:

1. Mention and context are represented in a common representation.
2. Only entities are remained in the anchor text.
3. Mention and entity are paired together and represented in a common representation, known as cross-encoder.
4. They use bi-encoders, which use two independent encoders (Wu et al., 2020) for entity and context, and generate entity embedding using these encoders.
5. They use bi-encoders and cross-encoders for processing mention, entity, and context.

B Features of Neural Entity Linking Models

In Table 4, we systematize the EL models in terms of their modifications and data requirements. Each column corresponds to a model feature.

- The **global** column shows whether a system uses a global solution (see Section 3.2.2).
- The **recognition** column refers to joint entity recognition and disambiguation models, where recognition and disambiguation of entities are performed collectively (Section 3.2.1).
- The **NIL prediction** column points out models that also label unlinkable mentions (Section 3.1.4).
- The **entity embedding** column presents how the entity disambiguation model uses entity representations, where *joint architecture* means the entity representations and the parameters of the disambiguation model are learned in the unified architecture, *separate architecture* indicates that representations are trained in one model and disambiguation parameters are learned in another, *pre-trained* denotes that the model uses pre-trained entity representations (Section 3.1.3).
- In the **candidate generation** column, the candidate generation methods are noted (Section 3.1.1). It contains the following options:
 - surface-form – simple surface-form matching heuristic;
 - dictionary – a dictionary with supplementary aliases for entities;
 - prior – filtering with precalculated mention-entity prior probabilities;
 - type classifier – Raiman and Raiman (2018) filter candidates using a classifier for an automatically learned type system;
 - tf-idf – Logeswaran et al. (2019) extract tf-idf scores based on entity description pages;
 - neighbours – the highest similar entities for the given mention are selected as candidates based on representations.
- The **zero-shot** column displays if an EL system provides a zero-shot approach (see Section 3.2.3).
- The **annotated text data** column shows whether a model uses an annotation or not. ‘In entity embedding’ denotes the models that do not use the annotated data for training, but the annotated data is used for training entity representation (see Section 3.2.3).
- The **cross-lingual** column refers to models, which provide cross-lingual EL solutions (Section 3.2.4).

Besides, the following superscript notations were used to denote specific features of methods:

1. In classification, the prior is checked by a threshold. This can be considered as a candidate selection step.
2. While training, they detect mentions, while testing, they assume mentions are detected.
3. They provide EL as a subsystem of language modeling.
4. They use document-level mention contexts while encoding.
5. Zero-shot in the sense of model adaptation to a new language using English annotated data, while the other zero-shot works solve the problem of model adaptation to a new domain without switching the language.

	Global	Recognition	NIL Prediction	Entity Embeddings	Candidate Generation	Zero-shot	Annotated Text Data	Cross-lingual
Sun et al. (2015)				joint architecture	surface-form prior		✓	
Francis-Landau et al. (2016)	✓			joint architecture	surface-form prior		✓	
Fang et al. (2016)	✓			separate architecture	prior ¹		✓	
Yamada et al. (2016)	✓			separate architecture	dictionary prior		✓	
Zwacklbauer et al. (2016)	✓		✓	separate architecture	surface-form prior neighbours		✓	
Tsai and Roth (2016)	✓		✓	separate architecture	prior		✓	✓
Pan et al. (2017)	✓		✓	separate architecture	dictionary		✓	✓
Cao et al. (2017)	✓			separate architecture	dictionary		in entity embedding	
Eshel et al. (2017)				pre-trained finetuned	dictionary		✓	
Ganea and Hofmann (2017)	✓			separate architecture	prior		✓	
Moreno et al. (2017)	✓		✓	separate architecture	surface-form		✓	
Gupta et al. (2017)	✓ ⁴			separate architecture	prior	✓		
Le and Titov (2018)	✓			pre-trained	prior		in entity embedding	
Newman-Griffis et al. (2018)				separate architecture	dictionary			
Radhakrishnan et al. (2018)	✓			separate architecture	dictionary		✓	
Kolitsas et al. (2018)	✓	✓		pre-trained	prior		✓	
Sil et al. (2018)	✓		✓	separate architecture	prior	✓ ⁶	✓	✓
Cao et al. (2018)	✓			pre-trained	prior		✓	
Raiman et al. (2018)	✓			n/a	prior type classifier		✓	✓
Shahbazi et al. (2019)				E-ELMo	prior		✓	
Logeswaran et al. (2019)		✓ ²		joint architecture	tf-idf	✓		
Gillick et al. (2019)				separate architecture	neighbours	✓	in entity embedding	
Peters et al. (2019) ³	✓	✓	✓	pre-trained	prior		in entity embedding	
Le and Titov (2019b)				joint architecture	surface-form			
Le and Titov (2019a)	✓			pre-trained	prior		in entity embedding	
Fang et al. (2019)	✓			pre-trained	dictionary		✓	
Martins et al. (2019)		✓	✓	pre-trained	dictionary		✓	
Yang et al. (2019)	✓			pre-trained	prior		✓	
Broscheit (2019)		✓		n/a	n/a		✓	
Wu et al. (2020)				joint architecture	neighbours	✓		
Yamada et al. (2020)	✓			joint architecture	prior		✓	

Table 4: Neural entity linking models are compared according to their features.