CROSS-LINGUAL ENTITY EXTRACTION AND LINKING FOR 300 LANGUAGES

BY

XIAOMAN PAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Dr. Heng Ji, Chair
Dr. Jiawei Han
Dr. Hanghang Tong
Dr. Kevin Knight

ABSTRACT

Information provided in languages which people can understand saves lives in crises. For example, language barrier was one of the main difficulties faced by humanitarian workers responding to the Ebola crisis in 2014. We propose to break language barriers by extracting information (e.g., entities) from a massive variety of languages and ground the information into an existing Knowledge Base (KB) which is accessible to a user in their own language (e.g., a reporter from the World Health Organization who speaks English only). The ambitious goal of this thesis is to develop a Cross-lingual Entity Extraction and Linking framework for 1,000 fine-grained entity types and 300 languages that exist in Wikipedia. Given a document in any of these languages, our framework is able to identify entity name mentions, assign a fine-grained type to each mention, and link it to an English KB if it is linkable.

Traditional entity linking methods rely on costly human annotated data to train supervised learning-to-rank models to select the best candidate entity for each mention. In contrast, we propose a novel unsupervised represent-and-compare approach that can accurately capture the semantic meaning representation of each mention, and directly compare its representation with the representation of each candidate entity in the target KB. First, we leverage a deep symbolic semantic representation the Abstract Meaning Representation [1] to represent contextual properties of mentions. Then we enrich the representation of each contextual word and entity mention with a novel distributed semantic representation based on cross-lingual joint entity and word embedding. We develop a novel method to generate cross-lingual data that is a mix of entities and contextual words based on Wikipedia. This distributed semantics enables effective entity extraction and linking. Because the joint entity and word embedding space is constructed across languages, we further extend it to all 300 Wikipedia languages and fine-grained entity extraction and linking for 1,000 entity types defined in YAGO [2]. Finally, using knowledge-driven question answering as a case study, we demonstrate the effectiveness of acquiring external knowledge using entity extraction and linking to improve downstream applications.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Prof. Heng Ji, for her advice, guidance, and support. She has introduced me to the Natural Language Processing field when I was just a junior undergraduate student. She has also provided me ample opportunities to fully engage me into this community. It is my great honor to work with Heng in the past seven years.

I want to thank my committee members: Prof. Jiawei Han, Prof. Hanghang Tong, and Dr. Kevin Knight, who have spent their valuable time on my thesis and provided me with constructive advice and comments for my research. Besides, I thank all my collaborators for their collaborations and inspirations.

I also thank my mentors: Dr. Yang Liu from Facebook, Dr. Dian Yu, and Dr. Chen Li from Tencent, for introducing me to the industrial community.

I would like to express my gratitude to my peers in the Blender Lab: Boliang Zhang, Di Lu, Tongtao Zhang, Lifu Huang, Ying Lin, Spencer Whitehead, Manling Li, and Qingyun Wang. I cherish their collaborations and support on research as well as their friendships.

I would also like to thank my best friend Feifei Pan and furry friends Goudan and Shizi, for their companion.

Finally, I thank my parents for their unconditional love, understanding, and support. They have always been supportive of all my decisions and encourage me to follow my dreams.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 MOTIVATIONS

For certain entities and events, a lot of new and detailed information is only available in documents written in a low-resource language for which there may be very few linguistic resources (annotated data, tools, etc.) available. For example, language barrier was one of the main difficulties faced by humanitarian workers responding to the Ebola crisis in 2014. When the Ebola outbreak started in 2014, news articles in Yoruba reported the newest updates with many details such as individual hospitals, researchers and local town names. In contrast, news articles in English mainly focused on general statistics such as the number of deaths, or non-local information such as a foreign government's reaction to the outbreak. Therefore, it will be highly valuable to automatically extract, link and fuse the knowledge across languages so we can construct a more complete profile in order to gain comprehensive understanding of an entity or event.

We propose to break language barriers by performing Cross-lingual Entity Extraction and Linking [3] for 300 languages that exist in Wikipedia. This task aims at automatically extracting and linking each named entity mention appearing in a *source* text document in any of the 300 languages to its unique entity referent in a *target* English knowledge base (KB), such as Wikipedia. For example, in the following Chinese sentence:

苹果是一家科技公司。(*Apple is a technology company.*)

Our framework will identify the entity name mention "苹果" (*Apple*), and link it to the entity `Apple_Inc.` in English Wikipedia.

A typical Cross-lingual Entity Extraction and Linking system works as follows. Entity mentions are extracted from source documents and translated into English if the source is in a foreign language. Given a mention m, the top N most likely entity referents from the KB are enumerated based on prior knowledge about which entities are most likely referred to using m. The candidate entities are re-ranked to ultimately link each mention to the top entity in its candidate list. Re-ranking consists of two key elements: *context representation* and *context comparison*. For a given mention, candidate entities are re-ranked based on a comparison of information obtained from the context of m with known structured and/or unstructured information associated with the top N KB entities, which can be considered as the "context" of the KB entity. The basic intuition is that the entity referent of m and related mentions should be similarly connected in the KB. Traditional Entity Linking methods use human annotated data to train supervised learning-to-rank model to select the best candidate

entity for each mention. The typical size of training data includes about 500 documents and 3000 entity mentions, which usually takes a team of expert annotators about half a year to prepare. Moreover, these approaches use hand-crafted features to represent the knowledge about an entity mention, including coarse-grained topic features [4, 5, 6, 7, 8, 9, 10, 11] and more fine-grained features such as contextual names, Wikipedia infobox, slot filling (entity profiling) and semantic categories [12, 13, 14, 15].

1.2 RESEARCH IN THIS THESIS

Symbolic Semantics based Unsupervised Approach. In stark contrast, in Chapter 2 we propose a novel unsupervised *represent-and-compare* paradigm for monolingual entity linking. The major challenge is that there might be many entity mentions in the context of a target entity mention that could potentially be leveraged for disambiguation. We will start by deriving contextual properties of each mention from Abstract Meaning Representation (AMR) [1] which includes more than 150 fine-grained semantic relation types, and thus these properties are discriminative enough to disambiguate entity mentions that current state-of-the-art systems cannot handle, without the need for entity linking training data. Specifically, for a given entity mention, we derive a rich symbolic context representation from AMR, facilitating the selection of an optimal set of *collaborator* entity mentions, i.e., those co-occurring mentions most useful for disambiguation. In previous approaches, collaborator sets tend to be too narrow or too broad, introducing noise. We then use unsupervised graph inference based on three simple measures for comparing each mention's context with each candidate entity's context in the KB. In addition, most state-of-the-art entity linking approaches rely on *collective inference*, where a set of *coherent mentions* are linked simultaneously by choosing an "optimal" or maximally "coherent" set of named entity targets - one target entity for each mention in the coherent set. We also propose an effective method based on AMR to partition all mentions in a document into coherent sets for collective linking.

Distributed Semantics based Unsupervised Approach. However, string matching based context comparison cannot capture homonyms or paraphrases, and it's limited to monolingual entity linking. We propose to further *decorate* each node in the mention's context graph with its continuous embedding representation learned from a large-scale data set based on distributied semantics, and each entity node in the KB with multi-hop knowledge graph embedding. Previous word embedding approaches have represented named entities as mere phrases, as a combination of the word vectors for each name component. This approach fails when names are rendered quite differently between languages and when the

mentions are ambiguous. For example, "*Ang Lee*" in English is "*Li An*" in Chinese. We propose a more principled treatment, pursuing a joint entity and word embedding across languages (Chapter 3). We propose a novel method to generate cross-lingual data that is a mix of entities and contextual words based on Wikipedia. We replace each anchor link in the source language with its corresponding entity title in English if it exists, or in the source language otherwise. A cross-lingual joint entity and word embedding learned from this kind of code-switched data not only can disambiguate linkable entities but can also effectively represent unlinkable entities. We adopt a linear mapping approach which leverages English entities as pivots to learn a rotation matrix and seamlessly align two embedding spaces into one. In this unified common space, multiple mentions are reliably disambiguated and grounded, which enables us to directly compute the semantic similarity between a mention in a source language and an entity in a target language (e.g., English), and thus we can perform cross-lingual entity linking in an unsupervised way.

Both representations alone have achieved results comparable with state-of-the-art supervised methods [16, 17, 18]. We propose to tightly integrate them in a nocel Graph Convolutional Network and extend the framework to all 300 Wikipedia languages and 1,000 entity types defined in YAGO [2].

Application on Knowledge-aware Question Answering. In Chapter 5, we explore entity extraction and linking methods for exploiting two sources of external knowledge for subject-area QA. The first enriches the original subject-area reference corpus with relevant text snippets extracted from an open-domain resource that cover potentially ambiguous entities in the question and answer options. As in other question answering research, the second method simply increases the amount of training data by appending additional in-domain subject-area instances. Experiments on three challenging multiple-choice science QA tasks demonstrate the effectiveness of our methods.

1.3 INNOVATIVE CLAIMS

The contributions of this thesis are summarized as follows:

- This is the first unsupervised method that exploits AMR for entity linking and achieves comparable performance with supervised methods. We show that AMR can better capture and represent the contexts of entity mentions than previous approaches.
- This is the first work to learn cross-lingual joint entity and word embedding in an unsupervised way and apply it to non-Wikipedia source documents.

- This is the first work that extends cross-lingual entity extraction and linking from several high-resource languages to 300 languages, from 7 main entity types to 1,000 fine-grained types.

1.4 RELATED PUBLICATIONS AND IMPACT

Some of the research work presented in this thesis has been published in the following papers. These papers have inspired the creation of the TAC-KBP2019 EDL task and TAC-KBP2020 REFUS task for thousands of fine-grained entity types. The programs and resources produced from this thesis thus far have been made publicly available for research purpose, and they have been widely used by the research community and demonstrated at several government simulated disaster monitoring exercises, and DARPA demo day, DARPA 60^{th} anniversary, and ARL transition day. These technologies have consistently achieved top performance in several international evaluations including NIST LoreHLT, TAC-KBP and TAC-SMKBP.

1.4.1 Core Publications

1. **Xiaoman Pan**, Taylor Cassidy, Ulf Hermjakob, Heng Ji and Kevin Knight, "*Unsupervised Entity Linking with Abstract Meaning Representation*". Proc. the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015).
2. **Xiaoman Pan**, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight and Heng Ji, "*Cross-lingual Name Tagging and Linking for 282 Languages*". Proc. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017).
3. **Xiaoman Pan**, Thamme Gowda, Heng Ji, Jonathan May and Scott Miller, "*Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Machine Translation*". Proc. EMNLP2019 Workshop on Deep Learning for Low-Resource Natural Language Processing.
4. **Xiaoman Pan***, Kai Sun*, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie and Dong Yu. "*Improving Question Answering with External Knowledge*". Proc. EMNLP2019 Workshop on Machine Reading for Question Answering.

1.4.2 Related Publications

1. Di Lu, **Xiaoman Pan**, Nima Pourdamghani, Shih-Fu Chang, Heng Ji and Kevin Knight. "*A Multi-media Approach to Cross-lingual Entity Knowledge Transfer*". Proc. the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016).
2. Ying Lin, **Xiaoman Pan**, Aliya Deri, Heng Ji and Kevin Knight. "*Leveraging Entity Linking and Related Language Projection to Improve Name Transliteration.* ". Proc. ACL2016 Workshop on Named Entities.
3. Lifu Huang, Jonathan May, **Xiaoman Pan**, Heng Ji, Xiang Ren, Jiawei Han, Lin Zhao and James Hendler. "*Liberal Entity Extraction: Rapid Construction of Fine-Grained Entity Typing Systems*". Big Data, Mar 2017, 5(1): 19-31.
4. Dian Yu, **Xiaoman Pan**, Boliang Zhang, Lifu Huang, Di Lu, Spencer Whitehead and Heng Ji. "*RPI_BLENDER TAC-KBP2016 System Description.*". Proc. Text Analysis Conference (TAC2016).
5. Boliang Zhang, Ying Lin, **Xiaoman Pan**, Di Lu, Jonathan May, Kevin Knight, Heng Ji. "*ELISA-EDL: A Cross-lingual Entity Extraction, Linking and Localization System*". Proc. the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT2018) Demo Track.
6. Heng Ji, **Xiaoman Pan**, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee and Cash Costello. "*Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking*". Proc. Text Analysis Conference (TAC2017).
7. Boliang Zhang, **Xiaoman Pan**, Ying Lin, Tongtao Zhang, Kevin Blissett, Samia Kazemi, Spencer Whitehead, Lifu Huang, Heng Ji, "*RPI BLENDER TAC-KBP2017 13 Languages EDL System*". Proc. Text Analysis Conference (TAC2017).
8. Mohamed Al-Badrashiny, Jason Bolton, Arun Tejavsi Chaganty, Kevin Clark, Craig Harman, Lifu Huang, Matthew Lamm, Jinhao Lei, Di Lu, **Xiaoman Pan**, Ashwin Paranjape, Ellie Pavlick, Haoruo Peng, Peng Qi, Pushpendre Rastogi, Abigail See, Kai Sun, Max Thomas, Chen-Tse Tsai, Hao Wu, Boliang Zhang, Chris Callison-Burch, Claire Cardie, Heng Ji, Christopher Manning, Smaranda Muresan, Owen C. Rambow, Dan Roth, Mark Sammons, Benjamin Van Durme, "*TinkerBell: Cross-lingual Cold-Start Knowledge Base Construction*". Proc. Text Analysis Conference (TAC2017).
9. Tongtao Zhang, Ananya Subburathinam, Ge Shi, Lifu Huang, Di Lu, **Xiaoman Pan**, Manling Li, Boliang Zhang, Qingyun Wang, Spencer Whitehead, Heng Ji, Alireza Zareian, Hassan Akbari, Brian Chen, Ruiqi Zhong, Steven Shao, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Dongyu Li, Xin Huang, Xujun Peng, Ryan Gabbard, Marjorie Freed-

man, Ali Sadeghian, Mayank Kejriwal, Ram Nevatia, Pedro Szekely, Ali Sadeghian and Daisy Zhe Wang. "*GAIA - A Multi-media Multi-lingual Knowledge Extraction and Hypothesis Generation System*". Proc. Text Analysis Conference (TAC2018).

10. Qingyun Wang, **Xiaoman Pan**, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji and Kevin Knight. "*Describing a Knowledge Base*". Proc. The 11th International Conference on Natural Language Generation.

CHAPTER 2: SYMBOLIC SEMANTICS BASED ENTITY LINKING

2.1 APPROACH OVERVIEW

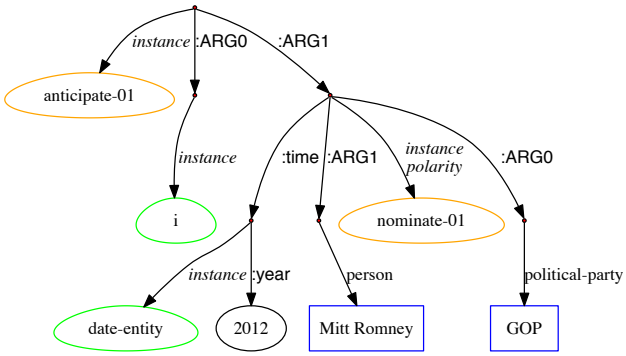
A typical Entity Linking system works as follows. Given a mention m (a string in a source document), the top N most likely entity referents from the KB are enumerated based on prior knowledge about which entities are most likely referred to using m. The candidate entities are re-ranked to ultimately link each mention to the top entity in its candidate list. Re-ranking consists of two key elements: *context representation* and *context comparison*. For a given mention, candidate entities are re-ranked based on a comparison of information obtained from the context of m with known structured and/or unstructured information associated with the top N KB entities, which can be considered the "context" of the KB entity. The basic intuition is that the entity referents of m and related mentions should be similarly connected in the KB. Next we will elaborate our approach to represent the context of each mention (Section 2.2) and each candidate entity (Section 2.3) respectively, and how to compare their contexts (Section 2.4).

2.2 MENTION REPRESENTATION BASED ON AMR

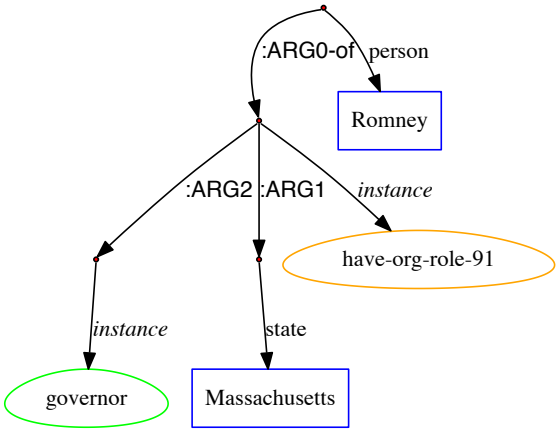
Abstract Meaning Representation (AMR) [1] is a sembanking language that captures whole sentence meanings in a rooted, directed, labeled, and (predominantly) acyclic graph structure. AMR utilizes multi-layer linguistic analysis such as PropBank frames, non-core semantic roles, coreference, named entity annotation, modality and negation to represent the semantic structure of a sentence. AMR strives for a more logical, less syntactic representation. Compared to traditional dependency parsing and semantic role labeling, the nodes in AMR are entities instead of words, and the edge types are much more fine-grained[1]. AMR thus captures deeper meaning compared with other representations more commonly used to represent mention context in Entity Linking.

We use AMR to represent semantic information about entity mentions expressed in their textual context. Specifically, given an entity mention m, we use a rule-based method to construct a Knowledge Network, which is a star-shaped graph with m at the hub, with leaf nodes obtained from entity mentions reachable by AMR graph traversal from m, as well as AMR node attributes such as entity type. A subset of the leaf nodes is selected as m's

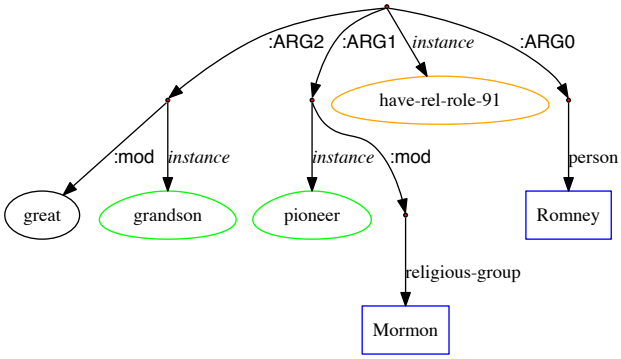
[1]AMR distinguishes between entities and concepts, the former being instances of the latter. We consider AMR concepts as entity mentions, and use AMR entity annotation for coreference resolution.



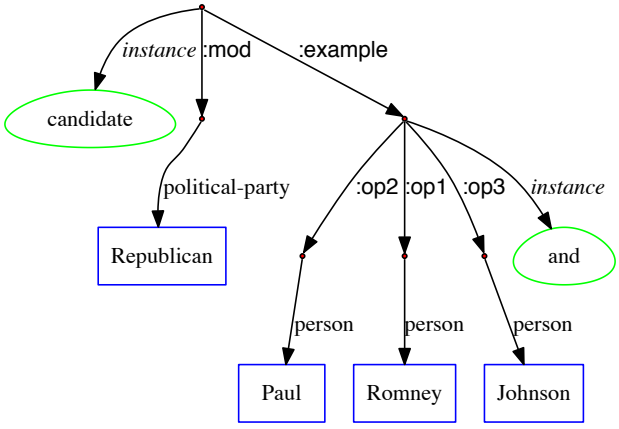
(a) *I am cautiously anticipating the GOP nominee in 2012 not to be* ***Mitt Romney****.*



(b) ***Romney*** *was the Governor of Massachusetts...*



(c) ***Romney*** *is the great-great-grandson of a Mormon pioneer...*



(d) *Republican candidates like* ***Romney****,* ***Paul****, and* ***Johnson****...*

Figure 2.1: AMR visualizations for the walk-through example.

collaborators using rules presented in the following subsections. Note that while we only evaluate linking of PER, ORG, and GPE entities, collaborators may be of any type. We also outline our efforts to use AMR to create sets of coherent entity mentions.

In each of the following subsections, we will describe the elements of AMR that are useful for context representation in EL. For each element, we will explain how our current system makes use of it (primarily, by using it to add entity mentions to a particular entity mention's set of collaborators). In doing so, we mainly refer to several examples from political discussion forums about "*Mitt Romney*", "*Ron Paul*" and "*Gary Johnson*". Their AMR graphs are depicted in Figure 2.1.

2.2.1 Entity Nodes

Each AMR node represents an entity mention, and contains its canonical name as inferred from the contextual sentence. This property is called *name expansion*. Consider the following sentence:

> "*Indonesia lies in a zone where the* ***Eurasian***, ***Philippine*** *and* ***Pacific*** *plates meet and occasionally shift, causing earthquakes and sometimes generating tsunamis.*"

Here, the nodes representing the three plates will be labeled as "*Eurasian Plate*", "*Philippine Plate*" and "*Pacific Plate*" respectively, even though these strings do not occur in the sentence. Note that these labels may be recovered primarily by appealing to syntactic reasoning, without consulting a KB. In our implementation, we consider these expanded names as mentions. These strings supersede raw mentions as input to the salience based candidate enumeration. Because the initial enumeration of entity candidates depends heavily on the mention's surface form, independent of context, name expansion will help us link "*Philippine*" to `en/Philippine_Sea_Plate` as opposed to the country.

An AMR node also contains an entity type. AMR defines 8 main entity types (Person, Organization, Location, Facility, Event, Product, Publication, Natural object, Other) and over one hundred fine-grained subtypes. For example, company, government organization, military, criminal organization, political party, school, university, research institute, team, and league are subtypes of organization. The fine-grained entity types defined in AMR help us restrict KB entity candidates for a given mention by encouraging entity type matching. For example, in

> "*The* ***Yuri dolgoruky*** *is the first in a series of new nuclear submarines to be commissioned this year but the Bulava nuclear-armed missile developed to equip the submarine has failed tests and the deployment prospects are uncertain.*"

AMR labels "*Yuri dolgoruky*" as a product instead of a person. We manually mapped AMR entity types to equivalent DBpedia types to inform type matching restrictions.

2.2.2 Semantic Roles

AMR defines core roles based on the OntoNotes [19] semantic role layer. Each predicate is associated with a sense and frame description. If a target entity mention m and a context entity mention n both play core roles for the same predicate, we consider n as a collaborator of m. Consider the following post:

"*Did **Palin** apologize to **Giffords**? He needs to conduct a beer summit between **Palin** and **NBC**.*"

We add "*Giffords*" and "*NBC*" as collaborators of "*Palin*", because they play core roles in both the *apologize-01* and *meet-03* events.

AMR defines new core semantic roles which do not exist in PropBank [20], NomBank [21], or Ontonotes [19]. Intuitively, the following special roles should provide discriminative collaborators:

- The ARG2 role of the *have-org-role-91* frame indicates the title held by an entity (ARG0), such as President and Governor, within a particular organization (ARG1).
- ARG2 and ARG3 of *have-rel-role-91* are used to describe two related entities of the same type, such as family members.

AMR defines a rich set of general semantic relations through non-core semantic roles. We choose the following subset of non-core roles to provide collaborators for entity mentions: *domain*, *mod*, *cause*, *concession*, *condition*, *consist-of*, *extent*, *part*, *purpose*, *degree*, *manner*, *medium*, *instrument*, *ord*, *poss*, *quant*, *subevent*, *subset*, *topic*.

2.2.3 Background Time and Location

AMR provides rich temporal and spatial information about entities and events. Types instantiated in AMR include time, year, month, day, source, destination, path and location. We exploit time and location entities as collaborators for entity mentions when they each play a role in the same predicate. For example, in the following post, the time role of the *die-01* event is "*2008*":

"*I just Read of **Clark**'s death in 2008.*"

We can link the mention "*Clark*" to the target entity `en/Arthur_C_Clark` in the KB, which contains the following triple:

```
< en/Arthur_C_Clark, date_of_death, 2018-03-19 >
```

Similarly, it's very challenging to link the abbreviation "*BMKG*", in the following sentence, to the correct target entity `en/Indonesian_Agency_for Meteorology,_Climatology_and_-Geophysics`, whose headquarters are listed as `en/Jakarta` in the KB:

"*It keeps on shaking. **Jakarta** BMKG spokesman Mujuhidin said*".

Here, "*Jakarta*" is added as a collaborator of "*BMKG*", since AMR labels it as the location of the organization, which facilitates the correct link because in the KB `en/Jakarta` is listed as its headquarter.

Authors often assume that readers will infer implicit temporal information about events. In fact, half of the events extracted by information extraction (IE) systems lack time arguments [22]. Therefore if an AMR parse includes no time information, we use the document creation time as an additional collaborator for mention in question. For example, knowing the document creation time "*2005-06-05*" can help us link "*Hsiung Feng*" in the following sentence

> "*The BBC reported that Taiwan has successfully test fired the* ***Hsiung Feng****, its first cruise missile*".

to `en/Hsiung_Feng_IIE`, which was deployed in 2005. Similarly, we include document creation location as a global collaborator.

2.2.4 Coreference

For linking purposes, we treat a coreferential chain of mentions as a single "mention". In doing so, the collaborator set for the entire chain is computed as the union over all of the chain's mentions' collaborator sets. From here on we refer to a coreferential chain of mentions as simply a "mention".

In order to construct a knowledge network across sentences, we use the following heuristic rules. If two names have a substring match (on a token-wise basis with stop words removed), or one name consists of the initials of another in all capital letters, then we mark them as coreferential. We replace all names in a coreferential chain with their canonical name, which may have been derived via name expansion: full names for people and abbreviations for organizations.

2.2.5 Knowledge Networks for Coherent Mentions

AMR defines a rich set of conjunction relations: *and*, *or*, *contrast-01*, *either*, *compared to*, *prep along with*, *neither*, *slash*, *between* and *both*. These relations are often expressed between entities that have other relations in common. We therefore group mentions connected by conjunction relations into sets of coherent mentions.

Figure 2.2 shows the expanded knowledge network that includes results from individual networks for each of the coherent mentions from the walk-through example. For each

coherent set, we merge the knowledge networks of all of its mentions[2].

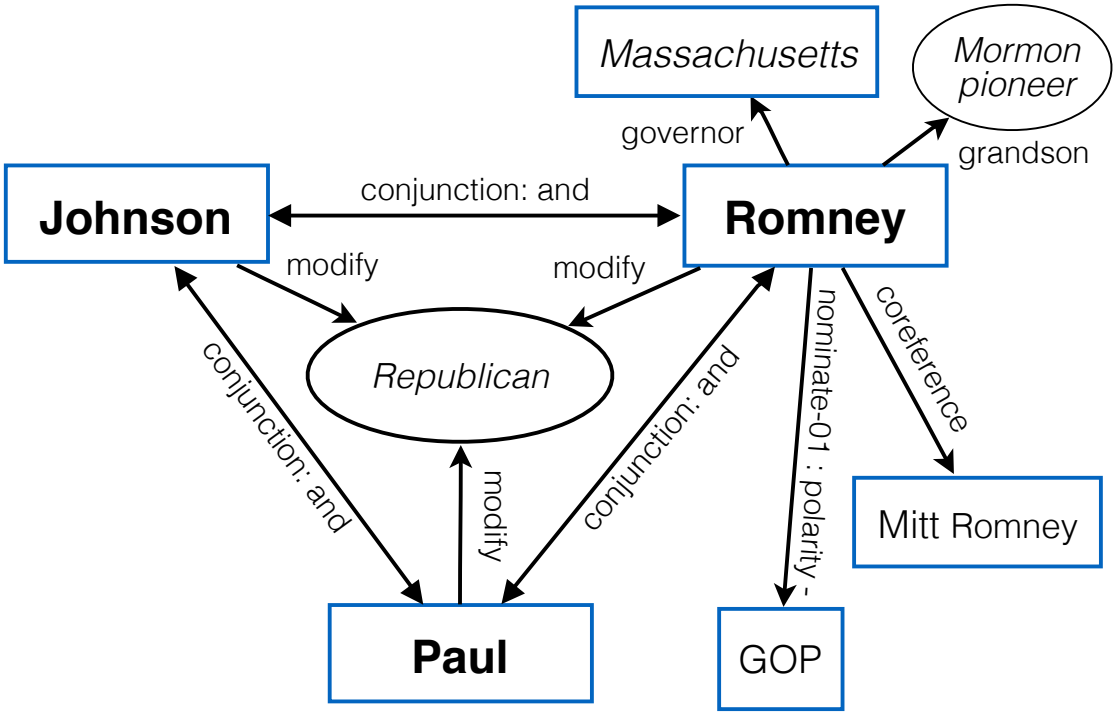


Figure 2.2: Knowledge network for mentions in source.

2.3 ENTITY REPRESENTATION IN KB

We combine Wikipedia with derivative resources to create the KB. The KB is a single knowledge network in which nodes are entities (Wikipedia articles) or constant values (e.g. a dollar amount or date), and the edges represent relations. We use this structure for context representation for entities, which together with context representation for mentions feeds re-ranking based on context comparison.

The KB is formally represented by triples:

```
< Entity_ID, Edge_Label, Node >
```

where `Entity_ID` is the entity's unique identifier, `Edge_Label` is relation type, and `Node` is the corresponding relation value - either another entity or a constant. These triples are derived from typed relations expressed within Wikipedia infoboxes, Templates, and Categories, untyped hyperlinks within Wikipedia article text, typed relations within DBpedia

[2]Recall that by mention, we mean a coreferential chain of mentions that may extend across sentences

and Freebase, and Google's "people also searched for" list[3]. Figure 2.3 shows a portion of the KB pertaining to the example in Figure 2.1.

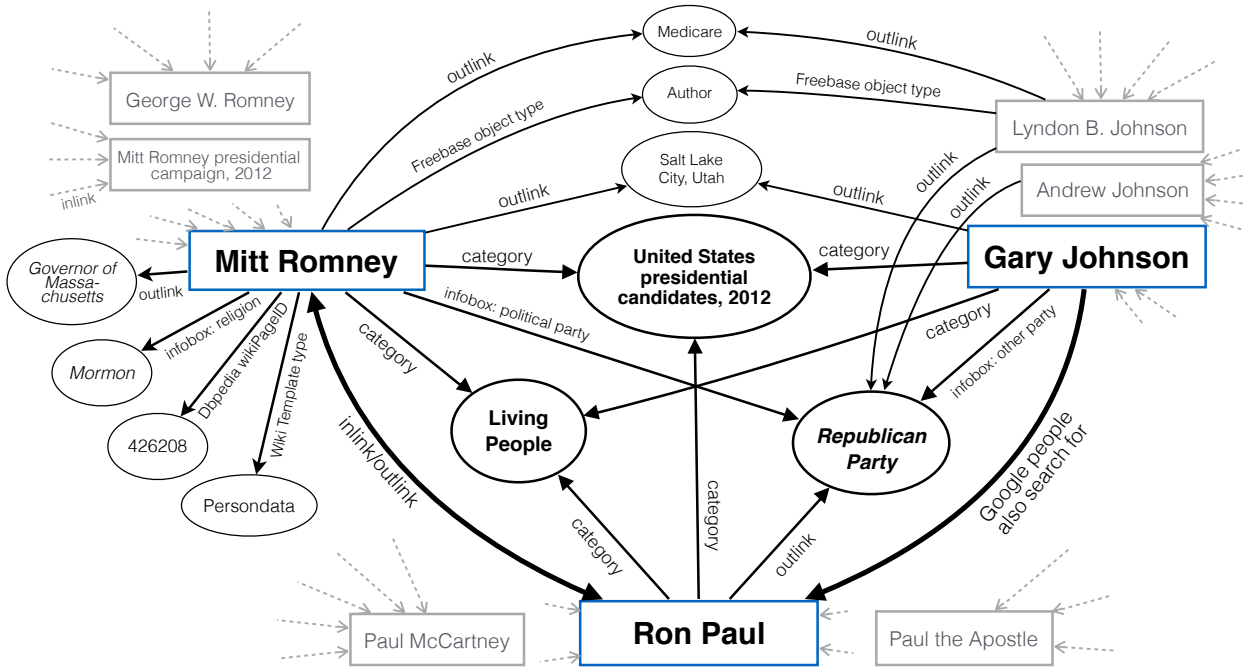


Figure 2.3: Knowledge network for entities in knowledge base.

In order to merge nodes from multiple KBs, we use the Wikipedia title as a primary key, and then use DBpedia *wikiPageID* and Freebase *Key* relations.

2.4 UNSUPERVISED ENTITY LINKING BASED ON MENTION AND ENTITY REPRESENTATION COMPARISON

In this section, we present our detailed algorithm to link each mention to a KB entity using a simple similarity measure over knowledge networks. Recall that a rule-based method has already been employed to construct star-shaped knowledge networks for individual mentions and entities (see sections 2.2 and 2.3; A KB knowledge network is the subnetwork of the entire KB centered at a candidate entity).

For each mention to be linked, an initial list of candidate entities are enumerated based on entity salience with respect to the mention, independent of mention context (Section 2.4.1)[4]. *Context collaborator* re-ranking proceeds in an unsupervised fashion agnostic to knowledge

[3]In response to a query entity Google provides a list of entities that "people also search for" - we add them into the entity's network.

[4]Here, "mention" means coreferential chain of mentions.

network edge labels using the Jaccard similarity measure computed between the mention and each entity, by taking their collaborator sets as inputs (Section 2.4.2).

We also describe *Context Coherence* re-ranking in terms of KB knowledge networks only, which constitutes preliminary steps toward unsupervised collective entity linking in section 2.4.3 based on the notion of coherence described in section 2.2.4. We leave a combination of the two re-ranking approaches to future work.

2.4.1 Salience

We use commonness [23] as a measure of context independent salience for each mention m, to generate an initial ranked list of candidate entities $E = (e_1, ..., e_N)$ where N is the cutoff for the number of candidates. In all experiments, we used N = 15 which can give us an oracle accuracy score 97.58%.

$$\text{Commonness}(m, e) = \frac{\text{count}(m, e)}{\sum_{e'} \text{count}(m, e')}$$

Here, count(m, e) is the number of hyperlinks with anchor text m and entity e within all of Wikipedia. As illustrated in Figure 2.3, using this salience measure "*Romney*" is successfully linked to `en/Mitt_Romney`. For the mention "*Paul*", the politician `en/Ron_Paul` is ranked at top 2 (less popular than the musician `en/Paul_McCartney`). For the mention "*Johnson*", the correct entity `en/Gary_Johnson` is ranked at top 9, after more popular entities such as `en/Lyndon_B._Johnson` and `en/Andrew_Johnson`.

2.4.2 Context Collaborator Based Re-ranking

Context collaborator based re-ranking is driven by the similarity between mention and entity knowledge networks. We construct knowledge network $g(m)$ for each mention m, and knowledge network $g(e_i)$ for each entity candidate e_i in m's entity candidate list E. We re-rank E according to Jaccard Similarity, which computes the similarity between $g(m)$ and $g(e_i)$

$$J(g(m), g(e_i)) = \frac{|g(m) \cap g(e_i)|}{|g(m) \cup g(e_i)|} \tag{2.1}$$

Note that the edge labels (e.g., *nominate-01* for a mention, or *infobox: religion* for an entity) are ignored, as the similarity metric operates over sets of collaborators (leaf nodes in the knowledge networks). For set intersection and union computation, elements are treated as

lists of lower-cased tokens with stop words removed, and two elements are considered equal if and only if they have one or more token in common. Due to the support from their neighbor `en/Republican` in the KB (Figure 2.3) which matches the neighbor "*Republican*" of mentions "*Paul*" and "*Johnson*" (Figure 2.2), `en/Ron_Paul` and `en/Gary_Johnson` are promoted to top 1 and top 3 respectively. `en/Gary_Johnson` is still behind two former U.S. presidents `en/Andrew_Johnson` and `en/Lyndon_B._Johnson` who also shares the neighbor `en/Republican` in the KB.

2.4.3 Context Coherence Based Re-ranking

Context coherence based re-ranking is driven by the similarity among KB entities. Let R_m be a set of coherent entity mentions, and R_E be the set of corresponding entity candidate lists, which are generated according to salience. Given R_E, we generate every combination of possible top candidate lists for the mentions in R_m, and denote the set of these combinations C_m. Formally, C_m is the Cartesian product of all candidate lists $E \in R_E$. In the walk-through example, R_m contains ["*Romney*", "*Paul*", "*Johnson*"], and C_m contains [`en/Mitt_Romney, en/Ron_Paul, en/Gary_Johnson`], [`en/Mitt_Romney, en/Paul_McCartney, en/Lyndon_Johnson`], etc. We compute coherence for each combination $c \in C_m$ as Jaccard Similarity, by applying a form of Equation 2.1 generalized to take any number of arguments to the set of knowledge networks for all entities in c, i.e., $\{g(e)|e \in c\}$.

The highest similarity combination is selected, yielding a top candidate for each $m \in R_m$. For example, compared to `en/Andrew_Johnson` and `en/Lyndon_Johnson`, `en/Gary_Johnson` is more coherently connected with `en/Mitt_Romney` and `en/Ron_Paul`, therefore it is promoted to top 1 with the coherence measure.

2.5 EXPERIMENTS

2.5.1 Data And Scoring Metric

For our experiments we use a publicly available AMR R3 corpus (LDC2013E117) that includes manual EL annotations for all entity mentions (LDC2014E15)[5]. For evaluation, we used all the discussion forum posts (DF), and news documents (News) that were sorted according to the alphabetic order of document IDs and taken as a tenth. The detailed data statistics are presented in Table 2.1.

[5]EL annotations are available to KBP shared task registrants (`nlp.cs.rpi.edu/kbp/2014`) via Linguistic Data Consortium (`www.ldc.upenn.edu`).

	PER	ORG	GPE	All
News	159	187	679	1,025
DF	235	129	224	588
All	394	316	903	1,613

Table 2.1: Total number of entity mentions in test set.

For each mention, we check whether the KB entity returned by an approach is correct or not. We compute accuracy for an approach as the proportion of mentions correctly linked.

2.5.2 Experiment Results

We focus primarily on context collaborator based re-ranking results. We compare our results with several baselines and state-of-the-art approaches in Table 2.2. In Table 2.3 we present preliminary results for collective linking.

Our unsupervised approach substantially outperforms the popularity based methods. More importantly, we see that AMR provides the best context representation for collaborator selection. Even system AMR outperforms not only baseline co-occurrence based collaborator selection methods but also outperforms the collaborator selection method based on human annotated core semantic roles which are used in traditional Semantic Role Labeling.

Figure 2.4 depicts accuracy increases as more AMR annotation is used in selecting collaborators. From the commonness baseline, additional knowledge about individual names leads to substantial gains followed by additional gains after incorporating links denoting semantic roles. Note that coreference here includes cross-sentence co-reference not based on AMR (Section 2.4.3). Furthermore, the results using human annotated AMR outperform the state-of-the-art supervised methods trained from a large scale EL training corpus, which rely on collective inference[6]. These results all verify the importance of incorporating a wider range of deep knowledge. Finally, Table 2.2 presents results in which our context coherence method is used where possible (i.e., those 215 mentions that are members of coherent sets according to our criteria as described in Section 2.4.3), and the context collaborator approach based on human AMR annotation is applied elsewhere.

Table 2.3 focuses on the 215 mentions that met our narrow criteria for forming a coherent set of mentions. We applied the context coherence based re-ranking method to collectively link those mentions. This approach substantially outperforms the co-occurrence baseline, and even outperforms the context collaborator approach applied to those 215 mentions,

[6]Note that the ground-truth EL annotation for the test set was created by correcting the output from supervised methods, so it may even favor these methods.

Approach		Definition	News	DF	Total
Popularity	Commonness	based on the popularity measure as described in Section 2.4.1.	89.8	69.0	82.2
	Google Search	use the top Wikipedia page returned by Google search using the mention as a keyword.	88.1	77.2	84.1
Supervised	State-of-the-art	supervised re-ranking using multi-level linguistic features for collaborators and collective inference, trained from 20,000 entity mentions from TAC-KBP2009-2014. We combined two systems [14, 24] using rules to highlight their strengths.	93.1	87.4	**91.0**
Unsupervised Context Collaborator Approach	Sen. Level Cooccurrence	sentence-level co-occurrence based collaborator selection	93.2	73.3	85.9
		(collaborators limited to human AMR-labeled named entities)	90.8	70.3	83.3
	Doc. Level Cooccurrence	document-level co-occurrence based collaborator selection	90.1	69.9	82.7
		(collaborators limited to human AMR-labeled named entities)	87.5	69.4	80.9
	Human AMR	using human annotated AMR nodes and edges.	93.6	86.9	**91.1**
	System AMR	using AMR nodes and edges automatically generated by an AMR parser [25].	90.2	85.7	**88.5**
	Human SRL	using human annotated core semantic roles defined in PropBank [20] and NomBank [21]: ARG0, ARG1, ARG2, ARG4 and ARG5.	93.3	71.2	85.2
Unsupervised Combined Approach	Human AMR	coherence approach used where possible (215 mentions), collaborator approach elsewhere (remaining 1398 mentions), using human annotated AMR nodes and edges.	94.3	88.3	**92.1**

Table 2.2: Accuracy (%) on test set (1613 mentions).

especially for discussion forum data.

2.5.3 Remaining Error Analysis and Discussion

A challenging source of errors pertains to the *knowledge gap* between the source text and KB. News and social media are source text genres that tend to focus on new information, trending topics, breaking events, or even mundane details about the entity. In contrast, the KB usually provides a snapshot summarizing only the entity's most representative and

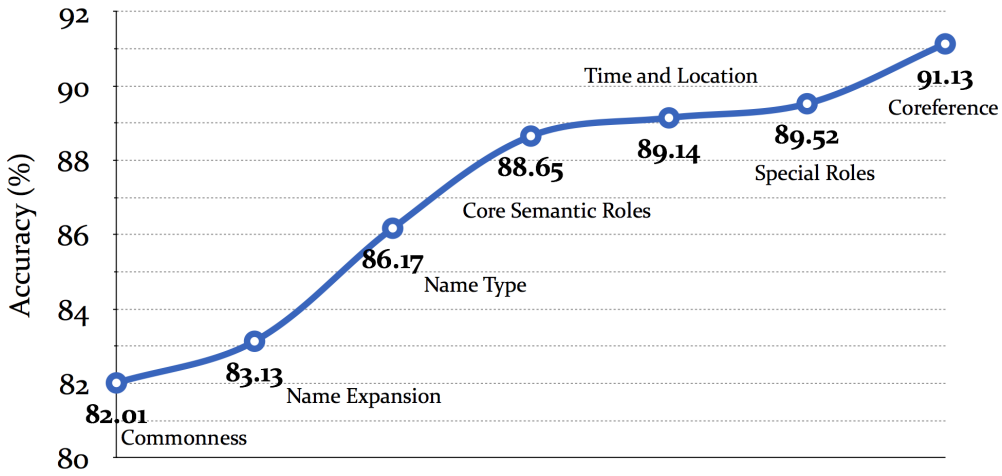


Figure 2.4: AMR annotation layers effects on accuracy.

Approach Description	News	DF	All
Coherence: coherence set built from within-sentence collaborators limited to human AMR-labeled Named Entities.	72.6	76.9	75.5
Coherence: coherence set built from human AMR conjunctions (Section 2.4.3)	96.7	95.2	**96.3**
Collaborator: used coherent set based on human AMR as collaborators.	91.5	82.3	88.8

Table 2.3: Context coherence accuracy (%) on 215 mentions which can form coherent sets.

important facts. A source-KB similarity driven approach alone will not suffice when a mention's context differs substantially from anything on the KB side. AMR annotation's synthesis of words and phrases from the surface texts into concepts only provides a first step toward bridging the knowledge gap. Successful linking may require (1) reasoning using general knowledge, or (2) retrieval of other sources that contain additional useful linking information. Table 2.4 illustrates two relevant examples that our system does not correctly link. In the first example, if we don't already know that Christie is the topic of discussion, as humans we might use our general knowledge that "governors veto bills" to pick the correct entity.

Using this type of knowledge presents interesting challenges (e.g., governors don't always veto bills, nor are they the only ones who can do so). In the second example, the rumor about this politician is not important enough to be reported in his Wikipedia page. We might first figure out, using cross-document coreference techniques, that a news article with the headline

"*Man Accused Of Making Threatening Phone Call To Kansas Gov.* ***Sam Brownback***

May Face Felony Charge..."

is talking about the same rumor. Then we might use biographical facts (e.g., Brownback is the governor of Kansas) from the article to enrich Brownback's knowledge network on the source side.

Type	Source	Knowledge Base
General Knowledge	**[Christies]$_{\mathbf{m}}$** denial of ***marriage*** privledges to ***gays*** will alienate independents and his "I wanted to have the people ***vote*** on it" will ring hollow.	**[Chris Christie]$_{\mathbf{e}}$** has said that he favoured New Jersey's law allowing ***same-sex*** couples to form civil unions, but would veto any bill legalizing ***same-sex marriage*** in New Jersey.
External Knowledge	Translation out of hype-speak: some kook made ***threatening*** noises at **[Brownback]$_{\mathbf{m}}$** and go ***arrested***.	**[Samuel Dale "Sam" Brownback]$_{\mathbf{e}}$** (born September 12, 1956) is an American politician, the 46th and current ***Governor*** of Kansas.

Table 2.4: Examples of knowledge gap.

Sometimes helpful neighbor concepts are omitted because the current collaborator selection criteria are too restricted. For example, "*armed*" and "*conflicts*" are informative words for linking "*The Stockholm Institute*" to `en/Stockholm_International_Peace_Research_Institute` in the following sentence

> "***The Stockholm Institute*** *stated that 23 of 25 major armed conflicts in the world in 2000 occurred in impoverished nations.*"

but they were not selected as context collaborators. In addition, our cross-sentence coreference resolution is currently limited to proper names. Expanding it to include nominals could further enrich context collaborators to overcome some remaining errors. For example, in the sentence,

> "*The first woman to serve on* ***SCOTUS***."

if we know "*The first woman*" is coreferential with "*Sandra Day O'Connor*" in the previous sentence, we can link "*SCOTUS*" to `en/Supreme_Court_of_the_United_States` instead of `en/Scotus_College`.

Finally, AMR parsers are only available for a few languages such as English and Chinese. To tackle these remaining challenges, in next chapter we propose to further enrich the representations of each mention, its contextual words, and each candidate entity using distributed semantics.

2.6 SUMMARY

Entity linking (EL) task requires a representation of the relations among entities in text. In this chapter, we show that the Abstract Meaning Representation (AMR) can better capture and represent the contexts of entity mentions for EL than previous approaches. We show that AMR enables EL performance comparable to the supervised state-of-the-art using an unsupervised, non-collective approach. In the future, this method can be further extended to combine collaborator and coherence methods into a unified approach, and to use edge labels in knowledge networks for context comparison (note that the last of these is quite challenging due to normalization, polysemy, and semantic distance issues). We have only applied a subset of AMR representations to the EL task, but we aim to explore how more AMR knowledge can be used for other more challenging Information Extraction and Knowledge Base Population tasks.

CHAPTER 3: ADDING DISTRIBUTED SEMANTICS FOR CROSS-LINGUAL ENTITY LINKING

3.1 APPROACH OVERVIEW

In the previous chapter, we have demonstrated the effectiveness of AMR on representing semantic structure of entity mentions and their textual contexts. However, matching symbolic representations highly relies on string match or heuristic rules, and thus fails to capture homonyms or paraphrases. For example, it is unlikely to match "*GOP*" with "*Republican Party*" using symbolic representations. In addition, the symbolic features are limited to sentence-level. In this chapter we propose to further enhance the framework by incorporating distributed semantic representations of mentions and entities. This hybrid representation can also serve as a bridge between languages and thus extend our framework from monolingual to cross-lingual entity linking.

The sheer amount of natural language data provides a great opportunity to represent named entity mentions by their probability distributions so that they can be exploited for many NLP applications. The distributional hypothesis [26] states that words often occurring in similar contexts tend to have similar meanings. However, named entity mentions are fundamentally different from common words or phrases in three aspects. First, the semantic meaning of a named entity mention (e.g., a person name "*Bill Gates*") is not a simple summation of the meanings of the words it contains ("*Bill*" + "*Gates*"). Table 3.1 shows the nearest neighbours for the words "*bill*", "*gates*", the summation "*bill*" + "*gates*", and the entity `Bill_Gates` using cosine similarity. We can see that the semantic meaning of the summation and the entity are very different, which indicates that named entity mentions need more complete representations. Second, entity mentions are often highly ambiguous

"*bill*"	"*gates*"	"*bill*" + "*gates*"	`Bill_Gates`
legislation	gate	bill	billionaire
senate-passed	doors	gates	ex-ceo
c-268	entrances	senate-passed	co-ceo
c-279	gateways	c-204	cto
bills	drawbridges	c-442	microsoft

Table 3.1: The nearest neighbours for the words "*bill*", "*gates*", the summation "*bill*" + "*gates*", and the entity `Bill_Gates` based cosine similarity.

in various local contexts. For example, "*Michael Jordan*" may refer to the basketball player or the computer science professor. Third, representing entity mentions as mere phrases

fail when names are rendered quite differently, especially when they appear across multiple languages. For example, "*Ang Lee*" in English is "*Li An*" in Chinese.

Fortunately, entities, the objects which mentions referring to, are unique and equivalent across languages. Many manually constructed entity-centric knowledge base resources such as Wikipedia are widely available. Even better, they are massively multilingual. For example, up to August 2018, Wikipedia contains 21 million inter-language links[1] between 302 languages. We propose a more principled treatment, pursuing a cross-lingual joint entity and word embedding based on multilingual Wikipedia.

Wikipedia contains rich entity anchor links. As shown in Figure 3.2, many mentions (e.g., "小米" (*Xiaomi*)) in a source language are linked to the entities in the same language that they refer to (e.g., `zh/`小米科技 (*Xiaomi Technology*)), and some mentions are further linked to their corresponding English entities (e.g., Chinese mention "苹果" (*Apple*) is linked to entity `en/Apple_Inc.` in English). We develop a simple yet effective approach to derive *code-switching* data by replacing each mention (anchor link) in the source language with its corresponding entity title in the target language if it exists, or in the source language otherwise. Using this kind of code-switching data, where each entity mention is treated as a unique disambiguated entity, we learn joint entity and word embedding representations for the source language and target language respectively.

Furthermore, we leverage these shared target language entities as pivots to learn a rotation matrix and seamlessly align two embedding spaces into one by linear mapping. In this unified common space, multiple mentions are reliably disambiguated and grounded, which enables us to directly compute the semantic similarity between a mention in a source language and an entity in a target language (e.g., English).

3.2 DATA GENERATION

Wikipedia contains rich entity anchor links. For example, in the following sentence from English Wikipedia:

> "**[[Apple Inc.|apple]]** is a technology company."

where **[[Apple Inc.|apple]]** is an anchor link that links the anchor text *"apple"* to the entity `en/Apple_Inc.` Traditional approaches to derive training data from Wikipedia usually replace each anchor link with its anchor text. These methods have two limitations: (1) Information loss: e.g., the anchor text *"apple"* itself does not convey information such as

[1] `https://en.wikipedia.org/wiki/Help:Interlanguage_links`

the entity is a company; (2) Ambiguity [27]: e.g., the fruit sense and the company sense of *"apple"* mistakenly share one surface form. Similar to previous work [28, 29, 30], we replace each anchor link with its corresponding entity title, and thus treat each entity title as a unique word. An example is illustrated in Figure 3.1. Using this kind of data mix of entity titles and contextual words, we can learn joint embedding of entities and words.

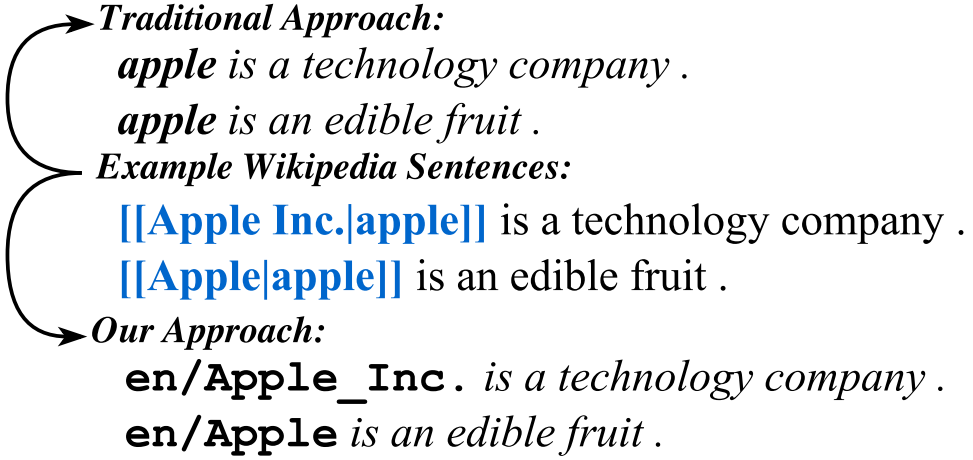


Figure 3.1: Comparison of data generated by the traditional approach and our approach on two example sentences from Wikipedia.

Moreover, the above approach can be easily extended to the cross-lingual setting by using Wikipedia inter-language links. We replace each anchor link in a source language with its corresponding entity title in a target language if it exists, and otherwise replace each anchor link with its corresponding entity title in the source language. An example is illustrated in Figure 3.2. Using the code-switching data generated from this approach, entities in a target

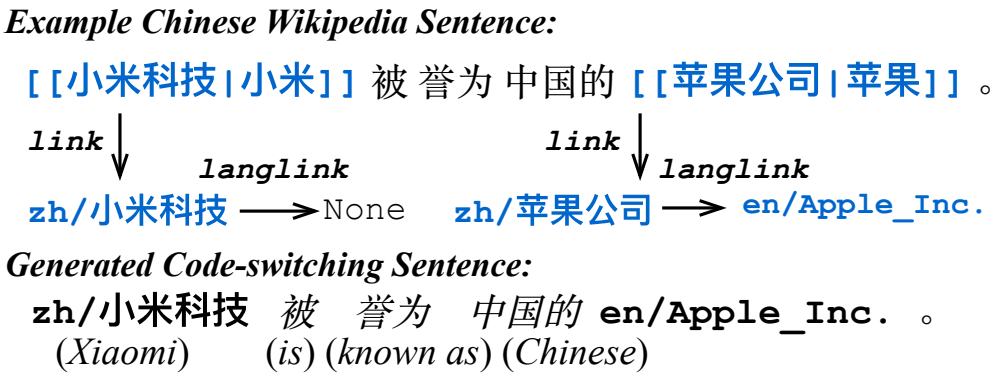


Figure 3.2: Using Wikipedia inter-language links to generate code-switched sentences which contain words and entities in a source language (e.g., Chinese) and entities in a target language (e.g., English).

language can be embedded along with words and entities in a source language, as illustrated in Figure 3.3. This joint representation has two advantages: (1) **Disambiguation**. For

Figure 3.3: Embedding which includes entities in English, and words and entities in Chinese (English words in brackets are human translations of Chinese words).

example, two entities `en/Apple_Inc.` and `en/Apple` can be differentiated by their distinct neighbors "电脑" (*computer*) and "水果" (*fruit*) respectively; (2) **Effective representation of unknown entities**. For example, the new entity `zh/`小米科技 (*Xiaomi Technology*), a Chinese mobile phone manufacturer, may not have an English Wikipedia page yet. But because it's close to neighbors such as `en/Microsoft`, "手机" (*phone*) and "公司" (*company*), we can infer it's likely to be a technology company.

3.3 LINEAR MAPPING ACROSS LANGUAGES

Word embedding spaces have similar geometric arrangements across languages [31]. Given two sets of independently trained word embedding, the source language embedding $\mathcal{Z}^S$ and the target language embedding $\mathcal{Z}^T$, and a set of pre-aligned word pairs, a linear mapping $\mathbf{W}$ is learned to transform $\mathcal{Z}^S$ into a shared space where the distance between the embedding of the source language word and the embedding of its pre-aligned target language word is minimized. For example, given a set of pre-aligned word pairs, we use $\mathbf{X}$ and $\mathbf{Y}$ to denote

two aligned matrices which contain the embedding of the pre-aligned words from $\mathcal{Z}^S$ and $\mathcal{Z}^T$ respectively. A linear mapping $\mathbf{W}$ can be learned such that

$$\underset{\mathbf{W}}{\text{argmin}}||\mathbf{WX} - \mathbf{Y}||_{\text{F}}$$

Previous work [32, 33] shows that enforcing an orthogonal constraint $\mathbf{W}$ yields better performance. Consequently, the above equation can be transferred to Orthogonal Procrustes problem [34] as

$$\underset{\mathbf{W}}{\text{argmin}}||\mathbf{WX} - \mathbf{Y}||_{\text{F}} = \mathbf{UV}^\top$$

Then $\mathbf{W}$ can be obtained from the singular value decomposition (SVD) of $\mathbf{YX}^\top$ such that

$$\mathbf{U\Sigma V}^\top = \text{SVD}(\mathbf{YX}^\top)$$

We propose using entities instead of pre-aligned words as anchors to learn such a linear mapping $\mathbf{W}$. The basic idea is illustrated in Figure 3.4.

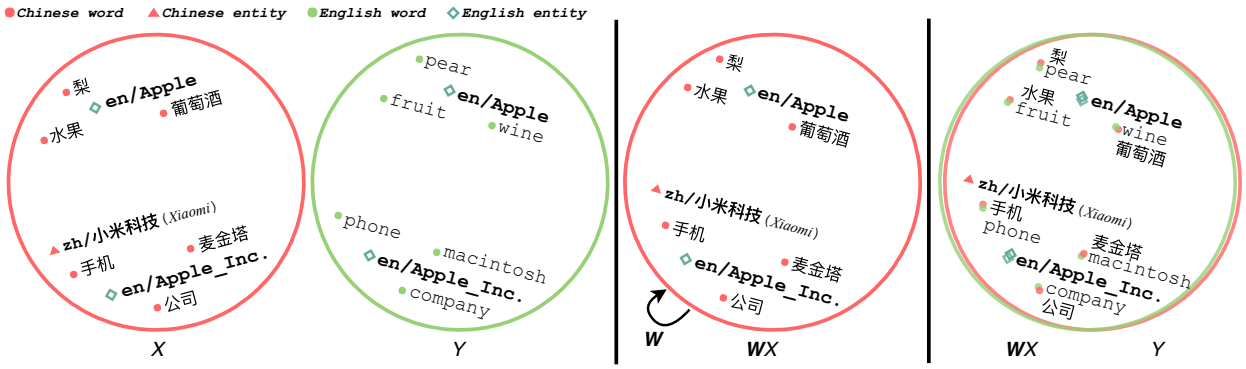


Figure 3.4: Using the aligned entities as anchors to learn a linear mapping (rotation matrix) which maps a source language embedding space to a target language embedding space.

We use $\mathcal{E}_T$ and $\mathcal{W}_T$ to denote the sets of entities and words in the target language associated with the target entity and word embedding as

$$\mathcal{Z}^T = \{\mathbf{z}^t_{e_1}, .., \mathbf{z}^t_{e_{|\mathcal{E}_T|}}, \mathbf{z}^t_{w_1}, .., \mathbf{z}^t_{w_{|\mathcal{W}_T|}}\}$$

Similarly, we use $\mathcal{E}_S$ and $\mathcal{W}_S$ to denote the sets of entities and words in the source language associated with the source entity and word embedding as

$$\mathcal{Z}^S = \{\mathbf{z}^s_{e_1}, .., \mathbf{z}^s_{e_{|\mathcal{E}_S|}}, \mathbf{z}^s_{w_1}, .., \mathbf{z}^s_{w_{|\mathcal{W}_S|}}\}$$

and use ${\mathcal{E}'}_T$ to denote the set of entities in the source language which are replaced with the corresponding entities in the target language, where ${\mathcal{E}'}_T \in \mathcal{E}_T$. Then $\mathcal{Z}^S$ can be represented as

$$\mathcal{Z}^S = \{\mathbf{z}^{t'}_{e_1}, .., \mathbf{z}^{t'}_{e_{|{\mathcal{E}'}_T|}}, \mathbf{z}^{s}_{e_1}, .., \mathbf{z}^{s}_{e_{|\mathcal{E}_S|-|{\mathcal{E}'}_T|}}, \mathbf{z}^{s}_{w_1}, .., \mathbf{z}^{s}_{w_{|\mathcal{W}_S|}}\}$$

Note that $\mathbf{z}^{t}_{e_i}$ and $\mathbf{z}^{t'}_{e_i}$ are the embedding of e_i in $\mathcal{Z}^T$ and $\mathcal{Z}^S$ respectively. Therefore, using entities in ${\mathcal{E}'}_T$ as anchors, we can learn a linear mapping $\mathbf{W}$ that maps $\mathcal{Z}^S$ into the vector space of $\mathcal{Z}^T$, and obtain the cross-lingual joint entity and word embedding $\mathcal{Z}$.

Moreover, we adopt the refinement procedure proposed by [34] to improve the quality of $\mathbf{W}$. A set of new high-quality anchors is generated to refine $\mathbf{W}$ learned from $\mathcal{E}'_T$. High-quality anchors refer to entities that are high frequency and entities that are mutual nearest neighbors. We iteratively apply this procedure to optimize $\mathbf{W}$. Specifically, at each iteration, the new high-quality anchors are exploited to learn a new mapping.

[34] also proposes the novel comparison metric, Cross-domain Similarity Local Scaling (CSLS), to relieve the hubness phenomenon, where some vectors (*hubs*) are the nearest neighbors of many others. For example, entity `en/United_States` is a *hub* in our proposed embedding space. By employing this metric, the similarity of isolated vectors is increased, while the similarity of vectors in dense areas is decreased. Specifically, given a mapped source embedding $\mathbf{Wx}$ and a target embedding $\mathbf{y}$, the mean cosine similarity of $\mathbf{Wx}$ and $\mathbf{y}$ for their K nearest neighbors in the other language, $r_T(\mathbf{Wx})$ and $r_S(\mathbf{y})$ are computed respectively. The comparison metric is defined as follow

$$\cos(\mathbf{Wx}, \mathbf{y}) - r_T(\mathbf{Wx}) - r_S(\mathbf{y})$$

[34] shows that the performance is essentially the same when $K = 5, 10, 50$. Following this work, we set $K = 10$.

3.4 CROSS-LINGUAL ENTITY LINKING BASED ON JOINT ENTITY AND WORD EMBEDDING

We use the above proposed embedding to implement the aforementioned similarity and coherence measures to directly compare the contexts of each mention and its candidate entity.

3.4.1 Similarity

Similarity refers to the context similarity between a mention and a candidate entity. Given a mention m, we consider the entire sentence containing m as its local context. Using our proposed cross-lingual joint entity and word embedding $\mathcal{Z}$, the vectors of context words are averaged to obtain the context vector representation of m:

$$\mathbf{v}_m = \frac{1}{|\mathcal{W}_m|} \sum_{w \in \mathcal{W}_m} \mathbf{z}_w$$

where $\mathcal{W}_m$ is the set of context words of m, and $\mathbf{z}_w \in \mathcal{Z}$ is the embedding of the context word w. We measure context similarity between m and each of its entity candidates by using the cosine similarity between $\mathbf{v}_m$ and entity embedding $\mathbf{z}_e \in \mathcal{Z}$ such that:

$$\mathcal{F}_{\text{txt}}(e) = \cos(\mathbf{v}_m, \mathbf{z}_e) = \frac{\mathbf{v}_m \cdot \mathbf{z}_e}{\|\mathbf{v}_m\| \, \|\mathbf{z}_e\|}$$

3.4.2 Coherence

Coherence is driven by the assumption that if multiple mentions appear together within a context window, their referent entities are more likely to be coherent in the KB. Previous work [4, 24, 35, 36, 37, 38, 39] considers the KB as a knowledge graph and models coherence based on the overlapped neighbors of two entities in the knowledge graph. These approaches heavily rely on explicit connections among entities in the knowledge graph and thus cannot capture the coherence between two entities that are implicitly connected. For example, two entities `en/Mosquito` and `en/Cockroach` only have very few overlapped neighbors in the knowledge graph, but they usually appear together and have similar contexts in text. Using our proposed embedding, the coherence score can be estimated by cosine similarity between the embedding of two entities. This coherence metric pays more attention to semantics.

We consider mentions that appear in the same sentence as coherent. Let m be a mention, and $\mathcal{C}_e$ be the set of corresponding entity candidates of m's coherent mentions. The coherence score for each of m's entity candidates is the average:

$$\mathcal{F}_{\text{coh}}(e) = \frac{1}{|\mathcal{C}_e|} \sum_{c_e \in \mathcal{C}_e} \cos(\mathbf{z}_e, \mathbf{z}_{c_e})$$

Finally, we linearly combine these two features with several other common mention disambiguation features as shown in Table 3.2.

Feature	Description
$\mathcal{F}_{\text{prior}}(e)$	Entity Prior: $\frac{\lvert A_{e,*}\rvert}{\lvert A_{*,*}\rvert}$, where $A_{e,*}$ is a set of anchor links that link to entity e and $A_{*,*}$ is all anchor links in the KB
$\mathcal{F}_{\text{prob}}(e\lvert m)$	Mention to Entity Probability: $\frac{\lvert A_{e,m}\rvert}{\lvert A_{*,m}\rvert}$, where $A_{*,m}$ is a set of anchor links with anchor text m and $A_{e,m}$ is a subset that links to entity e.
$\mathcal{F}_{\text{type}}(e\lvert m,t)$	Entity Type [39]: $\frac{p(e\lvert m)}{\sum_{e\mapsto t} p(e\lvert m)}$, where $e \mapsto t$ indicates that t is one of e's entity types. Conditional probability $p(e\lvert m)$ can be estimated by $\mathcal{F}_{\text{prob}}(e\lvert m)$.

Table 3.2: Mention disambiguation features.

3.5 EXPERIMENTS

3.5.1 Embedding Training

We use an April 1, 2018 Wikipedia dump to generate training data, and use the Skip-gram model in Word2Vec [40, 41] to learn the proposed embedding. The number of dimensions of the embedding is set to 300, and the minimal number of occurrences, the size of the context window, and the learning rate are set to 5, 5, and 0.025 respectively.

3.5.2 Linear Mapping

A large number of aligned entities can be obtained using the approach described in Section 3.2. For example, there are about 400,000 aligned entities between English and Spanish. However, the mapping algorithm does not align well if we try to align all anchors, for the reason that the embedding of rare entities is updated less often, and the context of rare entities is very likely to be distinct in each language. Since the mapping is linear, it is better to learn global mapping using only high-quality anchors. Therefore, only high-frequency entities are selected as anchors using the following salient metric

$$\text{prior}(e) = \frac{|A_{e,*}|}{|A_{*,*}|}$$

where $A_{e,*}$ is a set of anchor links that link to entity e and $A_{*,*}$ is all anchor links in the KB.

3.5.3 Entity Alignment

We divide all entity anchors into training and test sets, and the mapping quality is measured by alignment precision on the test set. We use 5,000 anchors for training and 1,500

anchors for testing for each language pair. Our proposed method is applied to 8 language pairs, with Table 3.3 showing the statistics and the performance. The results show that mapping a language to its related language (e.g., Ukrainian to Russian) usually achieves better performance.

Source-Target	**# of Training**	**# of Test**	$P@1$	$P@5$	$P@10$
Spanish-English	5,000	1,500	79.1	89.2	92.3
Italian-English	5,000	1,500	74.5	86.9	90.5
Russian-English	5,000	1,500	68.4	82.8	86.7
Turkish-English	5,000	1,500	59.0	79.9	86.3
Ukrainian-English	5,000	1,500	63.0	79.7	85.9
Chinese-English	5,000	1,500	63.1	83.8	89.2
Ukrainian-Russian	5,000	1,500	78.1	90.3	92.8
Russian-Ukrainian	5,000	1,500	75.8	90.2	93.7

Table 3.3: Linear mapping statistics and performance.

3.5.4 Parallel Sentence Mining

The second intrinsic evaluation we conduct is to mine parallel sentences from Wikipedia automatically using the proposed cross-lingual joint entity and word embedding $\mathcal{Z}$ (Section 3.3).

Wikipedia contributors tend to translate some content from existing articles in other languages while editing an article. Therefore, if there exists an inter-language link between two Wikipedia articles in different languages, they can be considered comparable and thus are very likely to contain parallel sentences. We represent a Wikipedia sentence in any language by aggregating the embedding of entities and words it contains by a weighted metric

$$\text{IDF}(t, \mathcal{S}) = \log\left(\frac{|\mathcal{S}|}{|\{s \in \mathcal{S} : t \in s\}|}\right)$$

where t is a term (entity or word), $\mathcal{S}$ is an article containing $|\mathcal{S}|$ sentences, and $|\{s \in \mathcal{S} : t \in s\}|$ is the total number of sentences containing t. The embedding of a sentence $\mathbf{v}_s$ can be computed as

$$\mathbf{v}_s = \frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} \text{IDF}(t, \mathcal{S}) \cdot \mathbf{z}_t$$

where $\mathcal{T}_s$ is the set of terms of s and $\mathbf{z}_t \in \mathcal{Z}$ is the embedding of t.

Given two comparable Wikipedia articles connected by an inter-language link, we compute the similarity of all possible sentence pairs using the CSLS metric described in Section 3.3 and rank them. If the CSLS score of a sentence pair is greater than a threshold (in this thesis, we set the threshold to 0.1), then the sentence pair is considered as parallel. An advantage of our approach is that it provides a similarity score for every term pair, which can be used for improving word alignment and entity alignment.

This approach can be applied to any two languages in Wikipedia. Therefore, we have mined parallel sentences from a total number of $\binom{302}{2}$ language pairs and made this data set publicly available for research purpose. Table 3.5 shows some examples of mined parallel sentences from Wikipedia, with word and entity alignment highlighted.

We randomly select 100 mined parallel sentence pairs for each of 3 language pairs, and ask linguistic experts to judge the quality of these sentence pairs (perfect, partial, or not parallel) and the accuracy of alignments. The results are shown in Table 3.4. We can see that the quality of mined parallel sentence is promising and the quality of word and entity alignment is decent.

Language Pairs	**Perfect**	**Partial**	**Word**	**Entity**
Chinese-English	81%	10%	92.3%	95.5%
Spanish-English	75%	13%	89.7%	91.1%
Russian-Ukrainian	70%	16%	82.4%	90.3%

Table 3.4: Quality of the mined parallel sentences (Perfect and Partial stand for the percentage of perfect and partial respectively; Word and Entity stand for the Accuracy of word and entity alignments respectively).

3.5.5 Cross-lingual Entity Linking

We compare our unsupervised approach to the top TAC-KBP2015 unsupervised system reported by [3]. In order to have a fair comparison with the state-of-the-art supervised methods, we also combine the features as described in Chapter 2 in a supervised fashion. More specifically, we generate all features on the training set of TAC15, and feed these features into a point-wise learning to rank algorithm Gradient Boosted Regression Trees [42]. The learning rate and the maximum depth of the decision trees are set to 0.01 and 4 respectively. A mention The F-score results are shown in Table 3.6, both of our unsupervised and supervised approaches significantly outperform the state-of-the-art methods.

We further observe that Context Similarity and Coherence features derived from our new joint entity and word embedding play significant roles. Without such features, the

Amharic - English
* ዓርብ የሳምንቱ ስድስተኛ ቀን ሲሆን ሐሙስ በኋላ ቅዳሜ በፊት ይገኛል ።
* Friday is the day after Thursday and the day before Saturday .
Yoruba - English
* Glasgow ni ilu totobijulo ni orile-ede Skotlandi ati eyi totobijulo keta ni Britani .
* Glasgow is the largest city in Scotland , and third largest in the United Kingdom .
Uyghur - English
* جۈمە ، پەيشەنبە بىلەن شەنبە ئوتتۇرسىدىكى ، ھەپتىنىڭ بەشىنچى كۈنىدۇر .
* Friday is the day after Thursday and the day before Saturday .
Vietnamese - English
* Bardolph là một làng thuộc quận McDonough , tiểu bang Illinois , HoaKỳ .
* Bardolph is a village in McDonough County , Illinois , United States .
Russian - Ukrainian
* Статья 2 - я Конституции СССР 1977 года провозглашала : « Вся власть в СССР принадлежит народу . * Стаття 2 - га Конституції СРСР 1977 року проголошувала : " Вся влада в СРСР належить народові . (*Article 2 of the Constitution of the USSR in 1977 proclaimed: "All power in the USSR belongs to the people."*)
Classical Chinese - Modern Chinese
* 至二战之时，南斯拉夫屡败，终为德意志、义大利所分。 * 在二次世界大战期间，南斯拉夫多次战败，分别被德国、意大利占领。 (*During the World War II, Yugoslavia was defeated several times and was occupied by Germany and Italy.*)

Table 3.5: Examples of mined parallel sentences from Wikipedia. A portion of alignments are highlighted using the same colors.

performance drops significantly, as shown in Table 3.6.

Moreover, we evaluate our approach on a collection constructed by [43] for 21 languages, which covers ground truth for the largest number of languages to date. Therefore we compare our approach with theirs that uses a supervised name transliteration model [44]. The entity linking results on non-NIL mentions are presented in Table 3.7. We can see that except Romanian, our approach outperforms or achieves comparable accuracy as their method on all languages, without using any additional resources or tools such as name transliteration.

Method	**English**	**Chinese**	**Spanish**
Best TAC15 Unsupervised	67.1	78.1	71.5
Our Unsupervised	**70.0**	**81.2**	**73.4**
w/o Context Similarity	66.9	79.0	70.6
w/o Coherence	68.5	78.6	71.4
Best TAC15 Supervised	73.7	83.1	80.4
[29]	-	83.6	80.9
Our Supervised	**74.8**	**84.2**	**82.1**
w/o Context Similarity	72.2	80.4	79.5
w/o Coherence	73.3	82.1	77.8

Table 3.6: F1 (%) on the evaluation set in TAC KBP 2015 Tri-lingual Entity Linking Track.

Language	# of Linkable Mentions	[43]	Our Approach
Arabic	661	70.6	**80.2**
Bulgarian	2,068	82.1	**84.1**
Chinese	956	-	91.0
Croatian	2,257	88.9	**90.8**
Czech	722	77.2	**85.9**
Danish	1,096	93.8	91.2
Dutch	1,087	92.4	89.2
Finnish	1,049	86.8	85.8
French	657	90.4	**92.1**
German	769	85.7	**89.7**
Greek	2,129	71.4	**79.8**
Italian	1,087	83.3	**85.6**
Macedonian	1,956	70.6	**71.6**
Portuguese	1,096	97.4	95.8
Romanian	2,368	93.5	88.7
Serbian	2,156	65.3	**81.2**
Spanish	743	87.3	**91.5**
Swedish	1,107	93.5	90.3
Turkish	2,169	92.5	92.2
Urdu	1,093	70.7	**73.2**

Table 3.7: Cross-lingual Entity Linking Performance for Low-resource Languages (accuracy %)

Note that [44] did not develop a model for Chinese even though Chinese data set was included in the collection.

3.6 COMBINING SYMBOLIC SEMANTICS AND DISTRIBUTED SEMANTICS

From the above we can see that entity mentions, contexts, and entities can be represented as vectors and these vectors can be directly used to compute similarity scores. However, unlike other NLP tasks, only discriminative contexts should be represented to disambiguate mentions in the Entity Linking task. Flat representations such as average embedding and language models may introduce some irrelevant information and lose attention to discriminative information.

For example, in the following sentence "***Parliament*** *approved plans in May 2001 to build an anti-drugs wall along the 925-kilometer border with Afghanistan.*", the contextual words such as "*approved plans*", "*build*", "*anti-drugs*", and "*wall*" are not discriminative enough to disambiguate entity mention "*Parliament*" because such contexts may be related to many different government organizations in the world. In contrast, the context "*border with Afghanistan*" match with the path in Wikidata that links two entities `Parliament_of_Pakistan` and `Afghanistan` (`Parliament_of_Pakistan` → *country* → `Pakistan` → *shares border with* → `Afghanistan`), and thus we can select `Parliament_of_Pakistan` as the correct entity.

Therefore, we propose to tightly integrate symbolic semantics and distributed semantics to better represent entity mentions and their contexts for entity linking. Compared to other traditional semantic parsing methods such as dependency parsing, AMR pays more attention to logic than syntax. For example, AMR does not annotate any individual words in a sentence, where the nodes in AMR are concepts instead of words. Besides, the edge types and entity types in AMR are much more fine-grained. Accordingly, AMR is more potent at capturing contextual properties that are discriminative enough to disambiguate entity mentions.

We combine symbolic semantics and distributed semantics in a Graph Convolutional Network (GCN) [45] to encode AMR graphs. The framework is illustrated in Figure 3.5. Given an input text sentence, we first apply an AMR parser to generate its AMR graph. To represent each word in the sentence, we concatenate its pre-trained word embedding proposed in Section 3.3, entity type embedding, and position embedding. We then feed the word sequence to a bi-directional long short term memory (Bi-LSTM) [46] network to encode the word order and generate the contextualized representation of each word. We assign a word representation to each node in the AMR graph, which are then used as the input for GCN. Given an AMR graph, we encode the graph contextual information following the previous work on event extraction [47]:

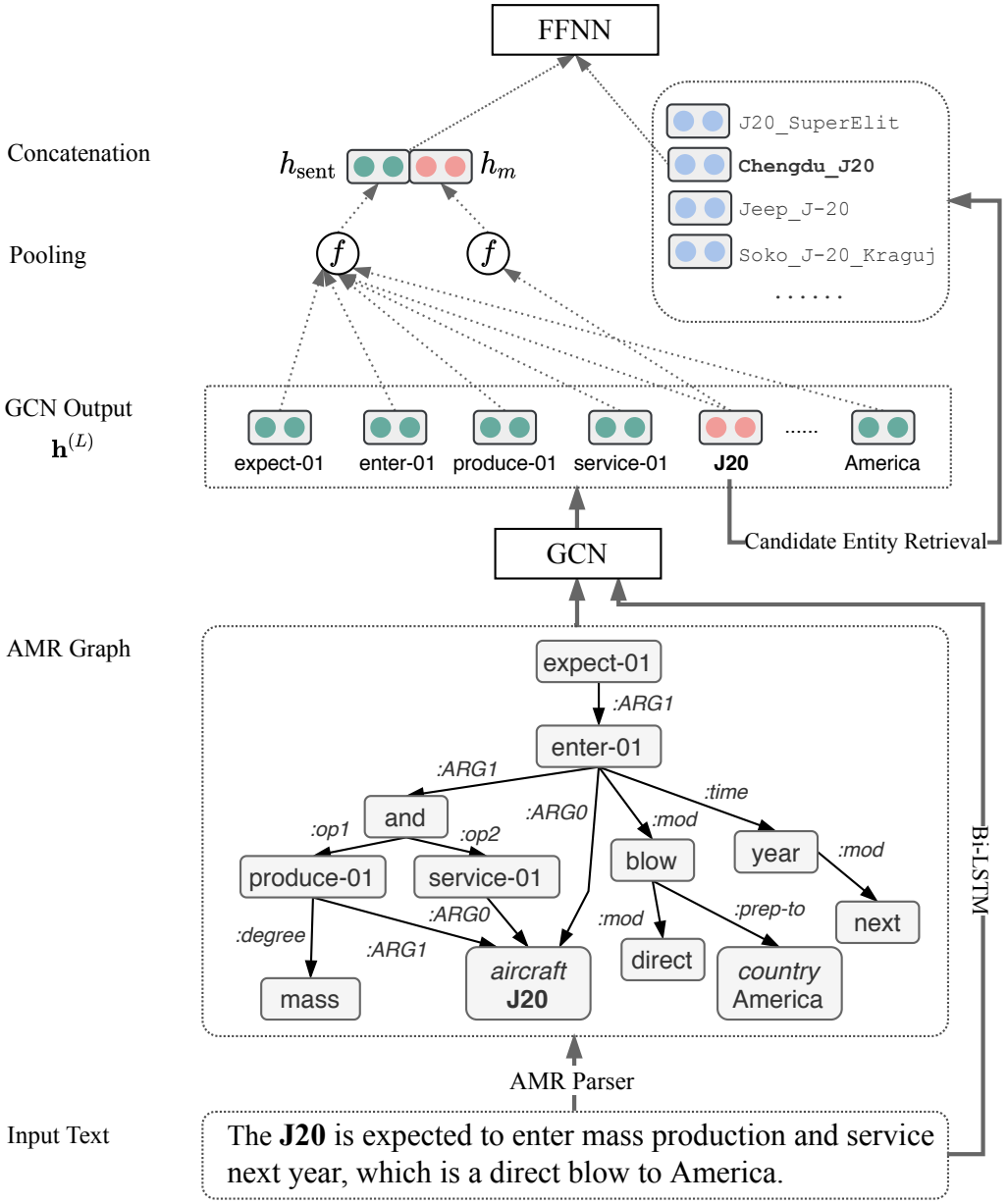


Figure 3.5: An illustration of the proposed framework to combine symbolic semantics and distributed semantics. To link entity mention "*J20*", we first generate its AMR graph using an AMR parser. Then we generate its sentence representation h_{sent} and entity mention representation h_m using contextualized GCN. We retrieve a list of candidate entities of "*J20*", and feed h_{sent}, h_m, and the entity embedding of each candidate through FFNN. The target entity (i.e., `Chengdu_J20`) is used as the positive training example, and the remaining candidate entities are used as the negative training examples.

$$w_i^{(k+1)} = S(\sum_{j \in \mathcal{N}(i)} g_{ij}^{(k)}(W_{E(i,j)} w_j^{(k)} + b_{E(i,j)}^{(k)})),$$

Method	Acc@1	Acc@5	Acc@10
Gold AMR			
Popularity	78.1	85.3	91.8
[16]	81.7	87.2	91.9
Our Method	**83.2**	**89.9**	**92.3**
System AMR [48]			
[16]	75.1	81.6	87.4
Our Method	**79.7**	**85.6**	**89.5**

Table 3.8: Accuracy (%) at K on LDC2019E81.

where $\mathcal{N}(i)$ is the set of neighbor nodes of w_i in the AMR graph, $E(i,j)$ is the edge type between w_i and w_j, g_{ij} is the gate following [47], k represents GCN layer number, S is the Sigmoid function, and W and b denote parameters of neural layers. We take the hidden states of the last GCN layer for each node and apply a max pooling function to generate sentence representation such that:

$$h_{\text{sent}} = f(\mathbf{h}^{(L)}) = f(\text{GCN}(\mathbf{h}^{(0)}))$$

where $\mathbf{h}^{(l)}$ denotes the hidden states at layer l of the GCN, and f is a max pooling function. We also obtain mention representation h_m from $\mathbf{h}^{(l)}$ as follows:

$$h_m = f(\mathbf{h}_{w_i}^{(L)})$$

where w_i is mention node.

We retrieve a list of candidate entities for each entity mention, where each candidate entity can be represented using the pre-trained entity embedding proposed in Section 3.3. We obtain the final representation by concatenating the sentence, mention and candidate entity representations, and feed them into a Feed-forward Neural Network (FFNN):

$$\Psi(m, e_j) = \text{FFNN}([h_{\text{sent}}; h_m; y_j])$$

where m denotes mention, e_j denotes candidate entity, y_j denotes the entity embedding of the candidate entity. The target entity is used as a positive training sample, and the remaining candidate entities are used as negative training samples. We minimize a loss function which enforces the scores of positive examples to be linearly separable from the scores of negative examples.

We conduct experiments on LDC2019E81 corpus which contains 193 documents and 6,153

entity mentions. We compare our method with a popularity baseline and a state-of-the-art method [16] in Table 3.8. Our method substantially outperforms the baseline and the state-of-the-art method, when using both human and system generated AMR graphs.

3.7 SUMMARY

In this chapter, we demonstrate a simple yet effective framework to learn cross-lingual joint entity and word embedding based on rich anchor links in Wikipedia. The learned embedding strongly enhances two downstream applications: cross-lingual entity linking and parallel sentence mining. The results have shown that our proposed method advances the state-of-the-art for unsupervised cross-lingual entity linking task. In the future, this framework can be extended to capture better representation of other types of knowledge elements such as relations and events.

CHAPTER 4: ENTITY EXTRACTION FOR 300 LANGUAGES AND 1,000 ENTITY TYPES

4.1 APPROACH OVERVIEW

Unlike entity linking, the most successful approaches for entity extraction are largely based on supervised learning models. However, training supervised models usually requires a massive amount of clean annotated data, which is often not available for low-resource languages and difficult to obtain during emergent settings. In order to compensate this data requirement, we propose a novel method to generate "silver-standard" entity annotations from Wikipedia markups and train a universal entity extraction system.

Wikipedia is a massively multi-lingual resource that currently hosts 300 languages and contains naturally annotated markups and rich knowledge structures through crowd-sourcing for 35 million articles in 3 billion words. Name mentions in Wikipedia are often labeled as anchor links to their corresponding referent pages. Each entry in Wikipedia is also mapped to external knowledge bases such as DBpedia[1], Wikidata[2], YAGO [2] and Freebase [49] that contain rich properties. Figure 4.1 shows an example of Wikipedia markups and KB properties.

The major challenges and our new solutions are summarized as follows.

Creating "Silver-standard" through cross-lingual entity transfer. The first step is to classify English Wikipedia entries into certain entity types and then propagate these labels to other languages. We exploit the English Abstract Meaning Representation (AMR) corpus [1] which includes both name tagging and linking annotations for fine-grained entity types to train an automatic classifier. Furthermore, we exploit each entry's properties in DBpedia as features and thus eliminate the need of language-specific features and resources such as part-of-speech tagging as in previous work.

Refine annotations through self-training. The initial annotations obtained from above are too incomplete and inconsistent. Previous work used name string match to propagate labels. In contrast, we apply self-training to label other mentions without links in Wikipedia articles even if they have different surface forms from the linked mentions.

Customize annotations through cross-lingual topic transfer. For the first time, we propose to customize name annotations for specific down-stream applications. Again, we use a cross-lingual knowledge transfer strategy to leverage the widely available English corpora to choose entities with specific Wikipedia topic categories Using the disaster scenario as a

[1] `http://wiki.dbpedia.org`

[2] `wikidata.org`

✤ **Wikipedia Article:**
Mao Zedong (d. 26 Aralık 1893 - ö. 9 Eylül 1976), Çinli devrimci ve siyasetçi. Çin Komünist Partisinin (ÇKP) ve Çin Halk Cumhuriyetinin kurucusu.

(*Mao Zedong (December 26, 1893 - September 9, 1976) is a Chinese revolutionary and politician. The founder of the Chinese Communist Party (CCP) and the People's Republic of China.*)

✤ **Wikipedia Markup:**
`[[Mao Zedong]] (d. [[26 Aralık]] [[1893]] - ö. [[9 Eylül]] [[1976]]), Çinli devrimci ve siyasetçi. [[Çin Komünist Partisi]]nin (ÇKP) ve [[Çin Halk Cumhuriyeti]]nin kurucusu.`

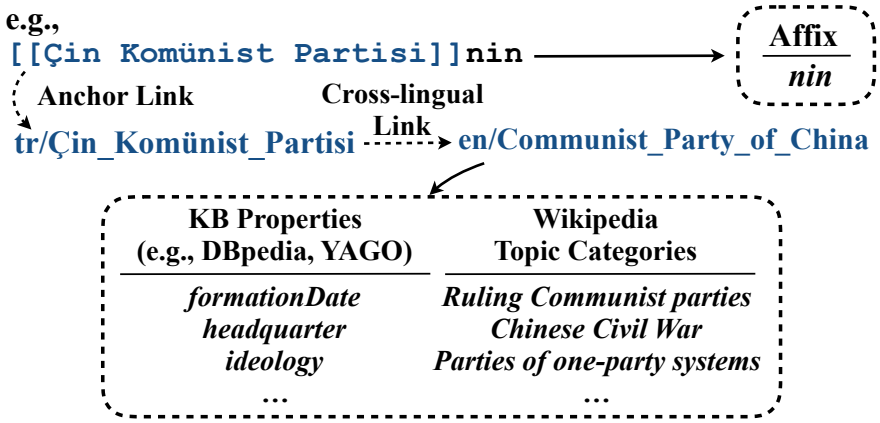


Figure 4.1: Examples of Wikipedia markups and KB properties.

use case, we apply English name tagging and linking on the Leidos corpus, which is a large collection of documents with disaster topics. Then we rank the Wikipedia topic categories of all linked entities based on their distributions in the Leidos corpus and select sentences including entities with top-ranked topic categories.

Derive morphology analysis from Wikipedia markups. Another unique challenge for morphologically rich languages is to segment each token into its stemming form and affixes. Previous methods relied on either high-cost supervised learning [50, 51, 52], or low-quality unsupervised learning [53, 54] which usually yields unsatisfactory performance due to the exclusion of language-specific knowledge. We exploit Wikipedia markups to automatically learn affixes as language-specific features for each language, and thereby eliminate the need to perform any deep language-specific morphological analysis.

4.2 ANNOTATION GENERATION

Our first step is to generate "silver-standard" name annotations from Wikipedia markups and train a universal name tagger. Figure 4.2 shows our overall procedure.

We will start by assigning an entity type or "other" to each English Wikipedia entry. We utilize the AMR corpus where each entity name mention is manually labeled as one of 139 types and linked to Wikipedia if it's linkable. In total we obtain 2,756 entity mentions, along with their AMR entity types, Wikipedia titles, YAGO entity types and DBpedia properties. For each pair of AMR entity type t^a and YAGO entity type t^y, we compute the Pointwise Mutual Information (PMI) [55] of mapping t^a to t^y across all mentions in the AMR corpus. Therefore, each name mention is also assigned a list of YAGO entity types, ranked by their PMI scores with AMR types. In this way, our framework produces three levels of entity typing schemas with different granularity: 4 main types (Person (PER), Organization (ORG), Geo-political Entity (GPE), Location (LOC)), 139 types in AMR, and 9,154 types in YAGO.

Then we leverage an entity's properties in DBpedia as features for assigning types. For example, an entity with a birth date is likely to be a person, while an entity with a population property is likely to be a geo-political entity. Using all DBpedia entity properties as features (60,231 in total), we train Maximum Entropy models to assign types with three levels of granularity to all English Wikipedia pages. In total we obtained 10 million English pages labeled as entities of interest.

[56] manually annotated 4,853 English Wikipedia pages with 6 coarse-grained types (Person, Organization, Location, Other, Non-Entity, Disambiguation Page). Using this data set for training and testing, we achieved 96.0% F-score on this initial step, slightly better than their results (94.6% F-score).

Next, we propagate the label of each English Wikipedia page to all of its corresponding entity mentions in all languages in the entire Wikipedia through mono-lingual redirect links and cross-lingual links. For example, after we successfully label the English page of "*Mitt Romney*" as `Politician|Person`, all entity mentions in all pages in other languages will be labeled as `Politician|Person`. After propagation, we use regular expression based sentence segmenter and word tokenizizerto separate sentences and tokens. For languages without spaces (e.g., Chinese), each character is considered as one token.

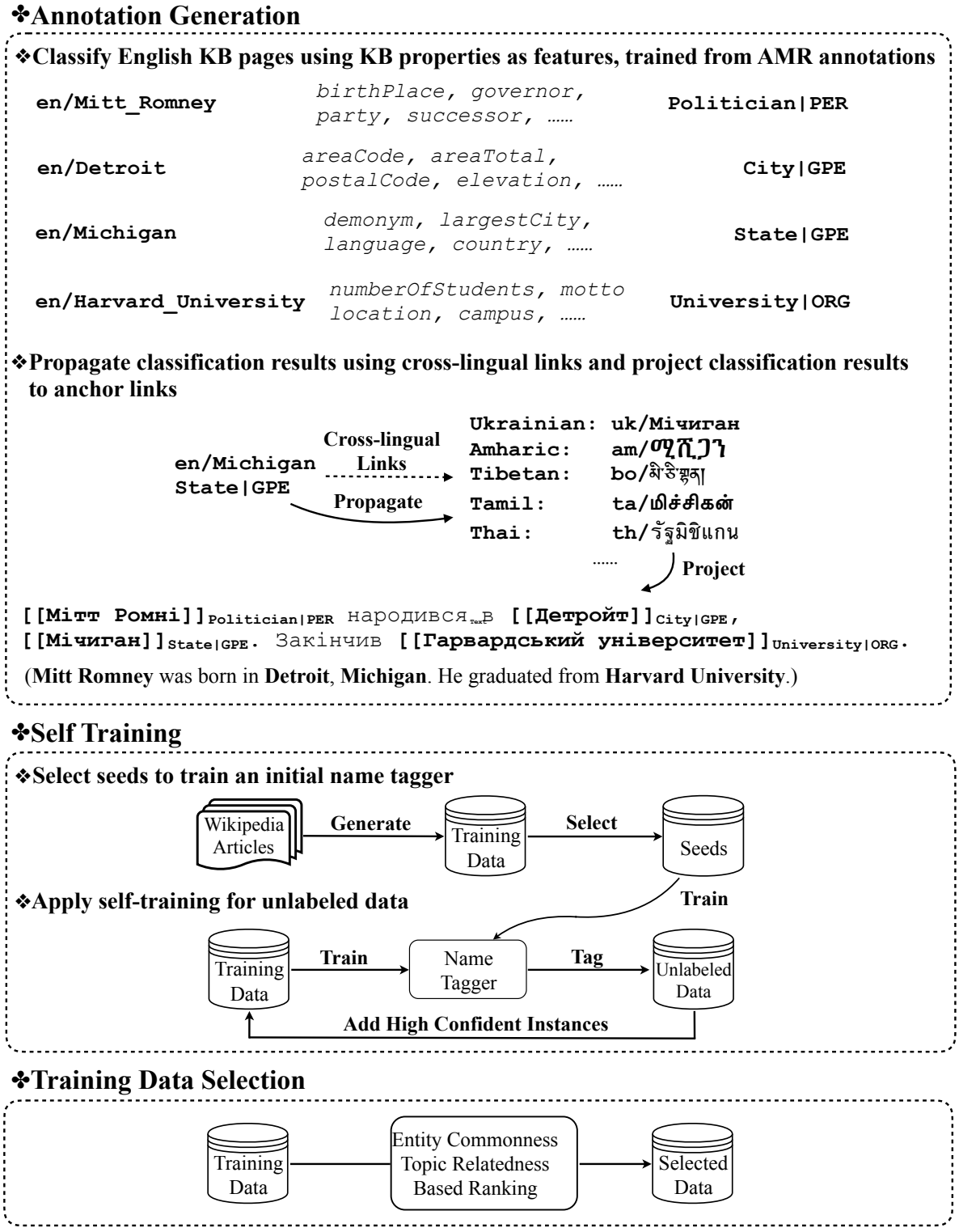


Figure 4.2: Entity Extraction annotation generation and training.

4.3 LEARNING MODEL AND KB DERIVED FEATURES

We will use a typical neural network architecture that consists of Bi-directional Long Short-Term Memory and Conditional Random Fields (CRFs) network [57] as our underlying learning model for the entity extraction system for each language. In the following we will describe how we acquire additional linguistic features to feed into the model.

When a Wikipedia user tries to link an entity mention in a sentence to an existing page, she/he will mark the title (the entity's canonical form, without affixes) within the mention using brackets "**[[ ]]**", from which we can naturally derive a word's stem and affixes for free.

For example, from the Wikipedia markups of the following Turkish sentence:

> Kıta Fransası, güneyde **[[Akdeniz]]den** kuzeyde [[Manş Denizi]] ve **[[Kuzey Denizi]]ne**, doğuda **[[Ren Nehri]]nden** batıda **[[Atlas Okyanusu]]na** kadar yayılan topraklarda yer alır. *(Metropolitan France extends from the Mediterranean Sea to the English Channel and the North Sea, and from the Rhine to the Atlantic Ocean.)*

we can learn the following suffixes: "*den*", "*ne*", "*nden*" and "*na*". We use such affix lists to perform basic word stemming, and use them as additional features to determine name boundary and type. For example, "*den*" is a noun suffix which indicates ablative case in Turkish. **[[Akdeniz]]den** means "*from Mediterranean Sea*". Note that this approach can only perform morphology analysis for words whose stem forms and affixes are directly concatenated.

4.3.1 Self-Training to Enrich and Refine Labels

The name annotations acquired from the above procedure are far from complete to compete with manually labeled gold-standard data. For example, if a name mention appears multiple times in a Wikipedia article, only the first mention is labeled with an anchor link. Therefore we propose to apply self-training to propagate and refine the labels.

We first train an initial entity extraction system using seeds selected from the labeled data. We adopt an idea from [58] which computes Normalized Pointwise Mutual Information (NPMI) [59] between a tag and a token

$$\text{NPMI}(\text{tag}, \text{token}) = \frac{\ln \frac{p(\text{tag},\text{token})}{p(\text{tag})p(\text{token})}}{-\ln p(\text{tag}, \text{token})}$$

[2]For languages that don't have word segmentation, we will consider each character as a token, and use character embeddings only.

Then we select the sentences in which all annotations satisfy NPMI(tag, token) $> \tau$ as seeds ($\tau = 0$ in our experiment).

For all Wikipedia articles in a language, we cluster the unlabeled sentences into n clusters ($n = 20$ in our experiment) by collecting sentences with low cross-entropy into the same cluster. Then we apply the initial tagger to the first unlabeled cluster, select the automatically labeled sentences with high confidence, add them back into the training data, and then re-train the tagger. This procedure is repeated n times until we scan through all unlabeled data.

4.3.2 Final Training Data Selection for Populous Languages

For some populous languages that have many millions of pages in Wikipedia, we obtain many sentences from self-training. In some emergent settings such as natural disasters it's important to train a system rapidly. Therefore we develop the following effective methods to rank and select high-quality annotated sentences.

Commonness: we prefer sentences that include common entities appearing frequently or linked by many other entities in Wikipedia. We rank names by their frequency and dynamically set the frequency threshold to select a list of common names. We first initialize the name frequency threshold S to 40. If the number of the sentences is more than a desired size D for training[3], we set the threshold $S = S + 5$, otherwise $S = S - 5$. We iteratively run the selection algorithm until the size of the training set reaches D for a certain S.

Topical Relatedness: various criteria should be adopted for different scenarios. Our previous work on event extraction [60] found that by carefully select $\frac{1}{3}$ topically related training documents for a test set, we can achieve the same performance as a model trained from the entire training set. Using an emergent disaster setting as a use case, we prefer sentences that include entities related to disaster related topics. We run an English name tagger [61] and entity linker [16] on the Leidos corpus released by the DARPA LORELEI program[4]. The Leidos corpus consists of documents related to various disaster topics. Based on the linked Wikipedia pages, we rank the frequency of Wikipedia categories and select the top 1% categories (4,035 in total) for our experiments. Some top-ranked topic labels include "*International medical and health organizations*", "*Human rights organizations*", "*International development agencies*", "*Western Asian countries*", "*Southeast African countries*"and "*People in public health*". Then we select the annotated sentences including names (e.g., "*World Health Organization*") in all languages labeled with these topic labels to train the

[3] D = 30,000 in our experiment.

[4] http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

final model. However, locations often dominate in natural disaster related documents, so this criteria tends to produce imbalanced entity type distribution. For example, the disaster related Cebuano sentences include 787,349 location names, significantly more than person names (1,251) and organization names (8,292). To address this issue, we adopt a simple down-sampling strategy as follows: if the mentions with the most dominant entity type is more than 500 times of the mentions with the least frequent type, we delete half of the sentences that only include mentions with the dominant entity type.

4.4 EXPERIMENTS ON COARSE-GRAINED TYPES

We first conduct experiments on three coarse-grained entity types: Person, Organization, Geo-political entity or Location.

4.4.1 Performance on Wikipedia Data

We conduct an evaluation using Wikipedia data as "silver-standard". For each language, we use 70% of the selected sentences for training and 30% for test. Figure 4.3 summarizes the performance, with some example languages marked for various ranges of data size.

Not surprisingly, entity extraction performs better for languages with more training mentions. The F-score is generally higher than 80% when there are more than 10K mentions, and it significantly drops when there are less than 250 mentions. The languages with low entity extraction performance can be categorized into three types: (1) the number of mentions is less than 2K, such as Atlantic-Congo (Wolof), Berber (Kabyle), Chadic (Hausa), Oceanic (Fijian), Hellenic (Greek), Igboid (Igbo), Mande (Bambara), Kartvelian (Georgian, Mingrelian), Timor-Babar (Tetum), Tupian (Guarani) and Iroquoian (Cherokee) language groups; Precision is generally higher than recall for most of these languages, because the small number of linked mentions is not enough to cover a wide variety of entities. (2) there is no space between words, including Chinese, Thai and Japanese. (3) they are not written in latin script, such as the Dravidian group (Tamil, Telugu, Kannada, Malayalam). The training instances for various entity types are quite imbalanced for some languages. For example, Latin data includes 11% PER names, 84% GPE/LOC names and 5% ORG names. As a result, the performance of ORG is the lowest, while GPE and LOC achieve higher than 75% F-scores for most languages.

Also note that since we don't have perfect annotations on Wikipedia data for any language, these results can be used to estimate how predictable our "silver-standard" data is, but they

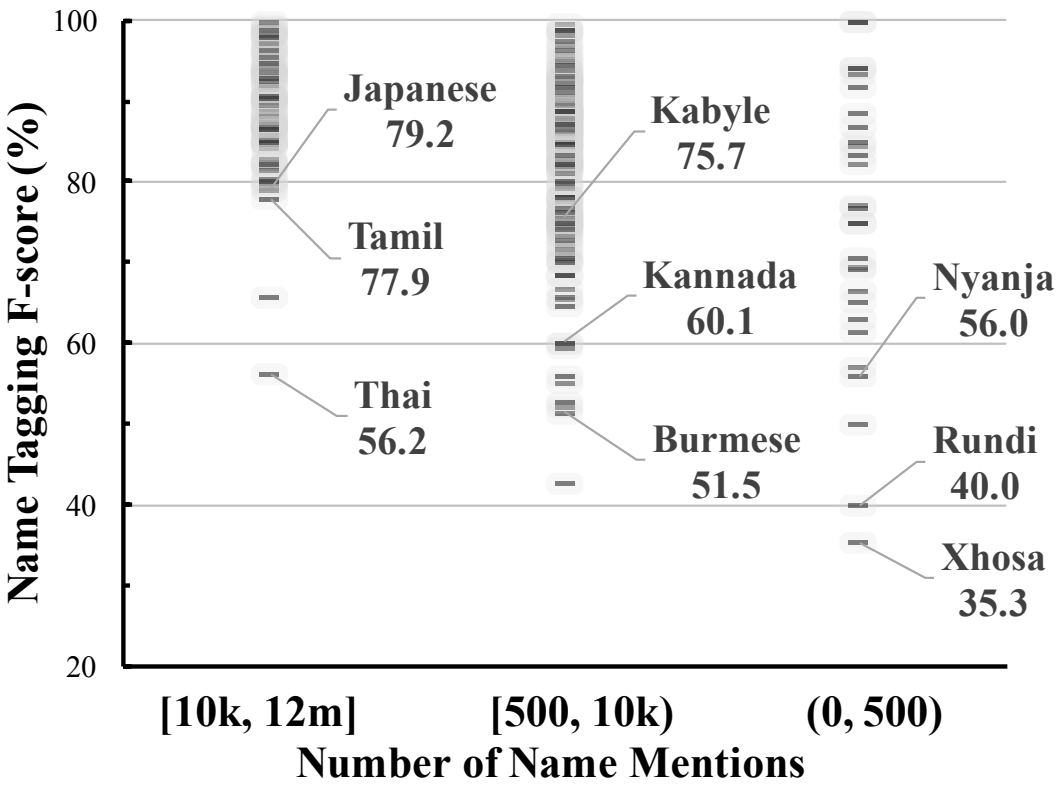


Figure 4.3: Summary of entity extraction F-score (%) on Wikipedia data.

are not directly comparable to traditional entity extraction results measured against gold-standard data annotated by human.

4.4.2 Performance on Non-Wikipedia Data

In order to have more direct comparison with state-of-the-art name entity extraction methods trained from human annotated gold-standard data, we conduct experiments on non-Wikipedia data in 9 languages for which we have human annotated ground truths from the DARPA LORELEI program. Table 4.1 shows the data statistics. The documents are from news sources and discussion fora.

The entity extraction results on LORELEI data set are presented in Table 4.2. We can see that our approach advances state-of-the-art language-independent methods [62, 63] on the same data sets for most languages, and achieves 6.5% - 17.6% lower F-scores than the models trained from manually annotated gold-standard documents that include thousands of name mentions. To fill in this gap, we would need to exploit more linguistic resources.

Language	*Gold* Training	*Silver* Training	Test
Bengali	8,760	22,093	3,495
Hungarian	3,414	34,022	1,320
Russian	2,751	35,764	1,213
Tamil	7,033	25,521	4,632
Tagalog	4,648	15,839	3,351
Turkish	3,067	37,058	2,172
Uzbek	3,137	64,242	2,056
Vietnamese	2,261	63,971	987
Yoruba	4,061	9,274	3,395

Table 4.1: Number of names in non-Wikipedia data.

Language	Training from *Gold*	Training from *Silver*	[62]	[63]
Bengali	61.6	**44.0**	34.8	43.3
Hungarian	63.9	47.9	-	-
Russian	61.8	49.4	-	-
Tamil	42.2	**35.7**	26.0	29.6
Tagalog	70.7	58.3	51.3	65.4
Turkish	66.0	**51.5**	43.6	47.1
Uzbek	56.0	44.2	-	-
Vietnamese	54.3	44.5	-	-
Yoruba	55.1	**37.6**	36.0	36.7

Table 4.2: Entity extraction F-score (%) on non-Wikipedia data.

4.5 EXPERIMENTS ON FINE-GRAINED TYPES

We further apply the same annotation generation framework described above and expand it to 1,000 salient fine-grained YAGO entity types.

4.5.1 Model

Unlike coarse-grained entity extraction, which only has few entity types, using sequence-to-sequence models with limited training data will not be effective for end-to-end fine-grained entity extraction. Therefore, we perform fine-grained entity typing on top of coarse-grained mention extraction results. We adopt an attentive classification model [64] that takes a mention with its contextual sentence and predicts the most possible fine-grained type for each mention. Unlike previous neural models that generally use fixed word embeddings and task-specific networks to encode each sentence, we employ contextualized word representa-

tions [65] that can capture word semantics in different contexts.

After that, we use a novel two-step attention mechanism to extract the most relevant information from the mention and its context as follows

$$\boldsymbol{m} = \sum_{M}^{i} a_i^m \boldsymbol{r}_i,$$

$$\boldsymbol{c} = \sum_{C}^{i} a_i^c \boldsymbol{r}_i,$$

where $\boldsymbol{r}_i \in \mathbb{R}^{d_r}$ is the vector of the i-th word, d_r is the dimension of $\boldsymbol{r}$, and attention scores a_i^m and a_i^c are calculated as

$$a_i^m = \text{Softmax}(\boldsymbol{v}^{m\top} \tanh\left(\boldsymbol{W}^m \boldsymbol{r}_i\right)),$$

$$a_i^c = \text{Softmax}(\boldsymbol{v}^{c\top} \tanh\left(\boldsymbol{W}^c (\boldsymbol{r}_i) \oplus \boldsymbol{m} \oplus p_i\right)),$$

$$p_i = \left(1 - \mu\Big(\min(|i-a|, |i-b|) - 1\Big)\right)^+,$$

where parameters $\boldsymbol{W}^m \in \mathbb{R}^{d_a \times d_r}$, $\boldsymbol{v}^m \in \mathbb{R}^{d_a}$, $\boldsymbol{W}^c \in \mathbb{R}^{d_a \times (2d_r+1)}$, and $\boldsymbol{v}^c \in \mathbb{R}^{d_a}$ are learned during training, a and b are indices of the first and last words of the mention, d_a is set to d_r, and μ is set to 0.1.

Next, we adopt a hybrid type classification model consisting of two classifiers. We first learn a matrix $\boldsymbol{W}^b \in \mathbb{R}^{d_t \times 2d_r}$ to predict type scores by

$$\tilde{y}^b = \boldsymbol{W}^b (m \oplus c),$$

where $\tilde{y}_i^b$ is the score for the i-th type.

We also learn to predict the latent type representation from the feature vector using

$$\boldsymbol{l} = \boldsymbol{V}^l (\boldsymbol{m} \oplus \boldsymbol{c}),$$

where $\boldsymbol{V}^l \in \mathbb{R}^{2d_r \times d_l}$. We then recover a type vector from this latent representation using

$$\tilde{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{l},$$

where $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ are obtained via Singular Value Decomposition (SVD) as

$$\boldsymbol{Y} \approx \tilde{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{L}^{\top},$$

where $\boldsymbol{U} \in \mathbb{R}^{d_t \times d_l}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d_l \times d_t}$, $\boldsymbol{L} \in \mathbb{R}^{N \times d_t}$, and $d_l \ll d_t$. Finally, we combine scores from two classifiers

$$\tilde{y} = \sigma(\boldsymbol{W}^b(\boldsymbol{m} \oplus \boldsymbol{c}) + \gamma \boldsymbol{W}^l \boldsymbol{l}),$$

where γ is set to 0.1. The training objective is to minimize the cross-entropy loss function as

$$J(\theta) = -\frac{1}{N}\sum_{i}^{N} \boldsymbol{y}_i \log \tilde{y}_i + (1 - \boldsymbol{y}_i) \log(1 - \tilde{\boldsymbol{y}}_i).$$

4.5.2 Performance on Wikipedia Data

Similar to coarse-grained entity extraction, we conduct both intrinsic and extrinsic experiments for fine-grained entity extraction. We first conduct experiments on 1,000 top frequent YAGO entity types using Wikipedia data. Table 4.3 and Table 4.4 show the data statistics and performance respectively.

Language	Training	Validation	Test
English	100,000	25,000	25,000
Chinese	100,000	25,000	25,000

Table 4.3: Number of names in Wikipedia data.

Language	Accuracy	Macro F1	Micro F1
English	46.6	74.1	75.0
Chinese	42.3	71.3	72.3

Table 4.4: Fine-grained entity typing results on Wikipedia data.

The following are some examples:

`Hormone`: ***Briain-derived neuotrophic factor*** *("**BDNF**"), another important gene in neural plasticity, has also been shown to have reduced methylation and increased transcription in animals that have undergone learning.*

`Infectious Disease`: *Notable exceptions include the **Large Pine Weevil**, which can kill young conifers.*

Dumpling: ***Shengjian mantou*** *is a type of small, pan-fried steamed buns which is a specialty of Shanghai.*

Lawsuit: *The landmark* ***Brown v. Board of Education*** *decision paved they way for PARC v. Commonwealth of Pennsylvania and Mills vs. Board of Education of District of Columbia, which challenged the segregation of students with special needs.*

Military Academy: *The year after, the prince went back to France, where he eventually entered the prestigious academy of* ***École spéciale militaire de Saint-Cyr-Coëtquidan****.*

Naval Gun: *She carried one* ***15 cm SK L/45 gun****, four* ***10.5 cm SK L/45 guns****, four* ***SK L/45 gun****, four* ***8.8 cm SK L/35 guns****, five* ***8.8 cm SK L/30 guns****, and one* ***8.8 cm SK L/30 gun*** *in a U-boat mounting.*

Vector: 一般的，令D 是作用于黎曼流形M 上的**向量丛** V 的一阶微分算子。*(In general, let D be the first-order differential operator of the* ***vector bundle*** *V acting on the Riemannian manifold M.)*

Footbridge: 而较高的一座哥特式塔楼于*1357* 年与**查理大桥** 一起由彼得帕尔莱勒兴建，直到*1464* 年才完成。*(The taller Gothic tower was built in 1357 by Peter Parleler with the* ***Charles Bridge*** *until 1464.)*

Automotive Technology: 同时，奥迪也在这一代*A4* 中引入了当时全新开发的*Tiptronic* **手自一体变速箱** *(At the same time, Audi also introduced a newly developed Tiptronic* ***tiptronic transmission*** *to this generation of A4 .)*

4.5.3 Performance on Non-Wikipedia Data

In order to evaluate the generalization capability of our approach, we conduct experiments on the TAC-SMKBP2020 fine-grained entity extraction task. The data set consists of documents in various genres such as newswire, newsgroup and discussion forum. We achieve 49.5% F-score, which is much 17% absolute higher than the second-best system.

4.6 SUMMARY

In this chapter, we demonstrate a framework that follows a fully automatic training and testing pipeline without any manual annotations or knowledge from native speakers. We

evaluate our framework on both Wikipedia articles and external formal and informal texts and obtained promising results. We further extend it to all 300 Wikipedia languages and fine-grained entity extraction for 1,000 entity types. Our multilingual entity extraction framework is applied to the largest number of languages to the best of our knowledge. In the future, we will explore the topological structure of related languages and exploit cross-lingual knowledge transfer to enhance the quality of extraction and linking. The general idea of deriving noisy annotations from KB properties can also be extended to other IE tasks such as relation extraction.

CHAPTER 5: APPLICATION ON KNOWLEDGE-AWARE QUESTION ANSWERING

Once we have an effective entity extraction and linking framework, we can apply it to many downstream knowledge-guided Natural Language Processing applications. We apply our framework to knowledge-aware question answering (QA), which requires both broad background knowledge and facts from the given subject-area reference corpus. We propose simple yet effective methods for exploiting two sources of external knowledge for subject-area QA. The first enriches the original subject-area reference corpus with relevant text snippets extracted from Wikipedia that cover potentially ambiguous entities in the question and answer options. As in other QA research, the second method simply increases the amount of training data by appending additional in-domain subject-area instances.

5.1 MOTIVATION

To answer questions relevant to a given text (e.g., a document or a book), human readers often rely on a certain amount of broad background knowledge obtained from sources outside of the text [66, 67]. It is perhaps not surprising then, that machine readers also require knowledge external to the text itself to perform well on question answering (QA) tasks.

We focus on multiple-choice QA tasks in subject areas such as science, in which facts from the given reference corpus (e.g., a textbook) need to be combined with broadly applicable external knowledge to select the correct answer from the available options [68, 69, 70]. For convenience, we call these **subject-area QA** tasks.

Question: a magnet will stick to ___?	
A. a belt buckle. ✓	**B**. a wooden table.
C. a plastic cup.	**D**. a paper plate.

Table 5.1: A sample problem from a multiple-choice QA task OpenBookQA [70] in a scientific domain (✓: correct answer option).

To correctly answer the question in Table 5.1, for example, scientific facts[1] from the provided reference corpus — {*"a magnet attracts magnetic metals through magnetism"* and *"iron is always magnetic"*}, as well as general world knowledge extracted from an external source such as {*"a belt buckle is often made of iron"* and *"iron is metal"*} are required. Thus,

[1]Ground truth facts are usually not provided in this kind of question answering tasks.

these QA tasks provide suitable testbeds for evaluating external knowledge exploitation and intergration.

Previous subject-area QA methods (e.g., [71, 72, 73]) explore many ways of exploiting structured knowledge. Recently, we have seen that the framework of fine-tuning a pre-trained language model (e.g., GPT [74] and BERT [75]) outperforms previous state-of-the-art methods [70, 76]. However, it is still not clear how to incorporate different sources of external knowledge, especially unstructured knowledge, into this powerful framework to further improve subject-area QA.

We investigate two sources of external knowledge (i.e., **open-domain** and **in-domain**), which have proven effective for other types of QA tasks, by incorporating them into a pre-trained language model during the **fine-tuning** stage. First, we identify entities in question and answer options and link these potentially ambiguous entities to an **open-domain** resource that provides unstructured background information relevant to the entities and used to enrich the original reference corpus (Section 5.2.2). In comparison to previous work (e.g., [77]), we perform information retrieval based on the enriched corpus instead of the original one to form a document for answering a question. Second, we increase the amount of training data by appending additional **in-domain** subject-area QA datasets (Section 5.2.3).

We conduct experiments on three challenging multiple-choice science QA tasks where existing methods stubbornly continue to exhibit performance gaps in comparison with humans: ARC-Easy, ARC-Challenge [68, 69], and OpenBookQA [70], which are collected from real-world science exams or carefully checked by experts. We fine-tune BERT [75] in a two-step fashion (Section 5.2.1). We treat entire Wikipedia as the **open-domain** external resource (Section 5.2.2) and all the evaluated science QA datasets (question-answer pairs and reference corpora) except the target one as **in-domain** external resources (Section 5.2.3). Experimental results show that we can obtain absolute gains in accuracy of up to 8.1%, 13.0%, and 12.8%, respectively, in comparison to the previous published state-of-the-art, demonstrating the effectiveness of our methods. We also analyze the gains and exposed limitations. While we observe consistent gains by introducing knowledge from Wikipedia, employing additional in-domain training data is not uniformly helpful: performance degrades when the added data exhibit a higher level of difficulty than the original training data (Section 5.3).

To the best of our knowledge, this is the first work to incorporate external knowledge into a pre-trained model for improving subject-area QA. Besides, our promising results emphasize the importance of external unstructured knowledge for subject-area QA. We expect there is still much scope for further improvements by exploiting more sources of external knowledge, and we hope the present empirical study can serve as a new starting point for researchers to identify the remaining challenges in this area.

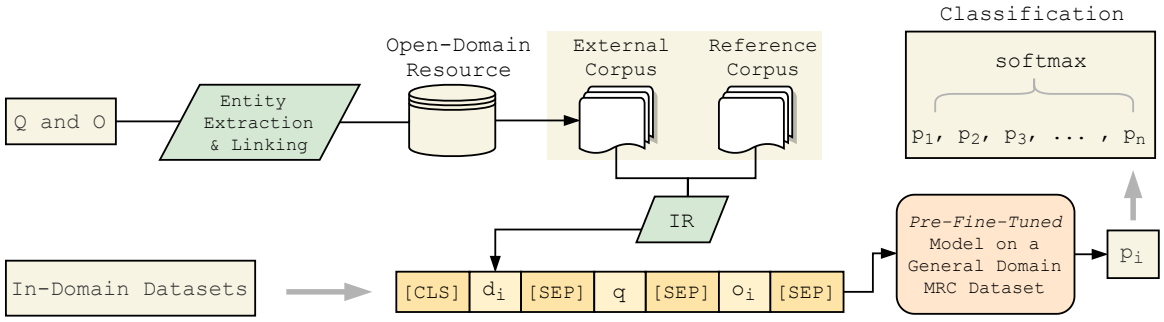


Figure 5.1: Overview of our framework (IR: information retrieval; MRC: machine reading comprehension). Q, O, q, o_i, d_i, and n denote the set of all questions, the set of all answer options, a question, one of the answer options associated with question q, the document (formed by retrieved sentences) associated with the (q, o_i) pair, and the number of answer options of q, respectively.

5.2 APPROACH OVERVIEW

In this section, we first introduce our BERT-based QA baseline (Section 5.2.1). Then, we present how we incorporate external open-domain (Section 5.2.2) and in-domain (Section 5.2.3) sources of knowledge into the baseline.

5.2.1 Baseline Framework

Given a question q, an answer option o_i, and a reference document d_i, we concatenate them with @ and # as the input sequence $@d_i\#q\#o_i\#$ to BERT [75], where @ and # stand for the classifier token `[CLS]` and sentence separator token `[SEP]` in BERT, respectively. A segmentation `A` embedding is added to every token before q (exclusive) and a segmentation `B` embedding to every other token, where `A` and `B` are learned during the language model pretraining of BERT. For each instance in the ARC (Easy and Challenge) and OpenBookQA tasks, we use Lucene [78] to retrieve up to top K sentences using the non-stop words in q and o_i as the query and then concatenate the retrieved sentences to form d_i [79]. The final prediction for each question is obtained by a linear plus softmax layer over the output of the final hidden state of the first token in each input sequence.

By default, we employ the following **two-step** fine-tuning approach unless explicitly specified. Following previous work [79] based on GPT [74], we first fine-tune BERT [75] on a large-scale multiple-choice machine reading comprehension dataset RACE [80] collected from English-as-a-foreign-language exams, which provides a ground truth reference document instead of a reference corpus for each question. Then, we further fine-tune the model on the

Question: Mercury, the planet nearest to the Sun, has extreme surface temperatures, ranging from 465°C in sunlight to −180°C in darkness. Why is there such a large range of temperatures on Mercury?
A. The planet is too small to hold heat.
B. The planet is heated on only one side.
C. The planet reflects heat from its dark side.
D. The planet lacks an atmosphere to hold heat. ✓

Table 5.2: A sample problem from the ARC-Challenge dataset [69] (✓: correct answer option).

target multiple-choice science QA datasets. For convenience, we call the model obtained after the first fine-tuning phase as a **pre-fine-tuned model**.

5.2.2 Utilization of External Knowledge from an Open-Domain Resource

Just as human readers activate their background knowledge related to the text materials [81], we link entities identified in questions and answer options to Wikipedia and provide machine readers with unstructured background information relevant to these entities, used to enrich the original reference corpus.

Entity Extraction and Linking: We first extract entity mentions from texts. Most mention extraction systems (e.g., [82]) are trained using pre-defined classes in general domain such as PERSON, LOCATION, and ORGANIZATION. However, in ARC and OpenBookQA, the vast majority of mentions are from scientific domains (e.g., *"rotation"*, *"revolution"*, *"magnet"*, and *"iron"*). Therefore, we simply consider all noun phrases as candidate entity mentions, which are extracted by a noun phrase chunker. For example, in the sample problem in Table 5.2, we extract entity mentions such as *"Mercury"*.

Then each entity mention is disambiguated and linked to its corresponding entity (page) in Wikipedia. For example, the ambiguous entity mention *"Mercury"* in Table 5.2 should be linked to the entity `Mercury_(planet)` rather than `Mercury_(element)` in Wikipedia. For entity disambiguation and linking, we simply adopt an existing unsupervised approach [16] that first selects high quality sets of entity *collaborators* to feed a simple similarity measure (i.e., Jaccard) to link entity mentions.

Reference Corpus Enrichment: We apply entity extraction and linking to the text of all questions and answer options. Then, for each linked entity, we extract Wikipedia sentences that contain this entity and all sentences from the Wikipedia article of this entity without

removing redundant information. For example, the following sentence in the Wikipedia article of `Mercury_(planet)` is extracted: *"Having almost no **atmosphere** to retain **heat**, it has surface temperatures that vary diurnally more than on any other planet in the Solar System."*, which can serve as a reliable piece of evidence to infer the correct answer option D for the question in Table 5.2.

Most previous methods [76, 77, 83, 84] perform information retrieval on the reference corpus to retrieve relevant sentences to form reference documents. In contrast, we retrieve relevant sentences from the **combination** of an open-domain resource and the original reference corpus to generate a reference document for each (question, answer option) pair. We still keep **up to top** K sentences for each reference document (Section 5.2.1). See the framework overview in Figure 5.1.

5.2.3 Utilization of External Knowledge from In-Domain Data

Since there are a relatively small number of training instances available for a single subject-area QA task (see Table 5.3), instead of fine-tuning a pre-fine-tuned model on a single target dataset, we also investigate into fine-tuning a pre-fine-tuned model on multiple in-domain datasets simultaneously. For example, when we train a model for ARC-Challenge, we use the training set of ARC-Challenge together with the training, development, and test sets of ARC-Easy and OpenBookQA. We also explore two settings with and without merging the reference corpora from different tasks. We introduce more details and discussions in Section 5.3.2 and Section 5.3.6.

5.3 EXPERIMENTS

5.3.1 Datasets

In our experiment, we use RACE [80] — the largest existing multiple-choice machine reading comprehension dataset collected from real and practical **language** exams — in the pre-fine-tuning stage. Questions in RACE focus on evaluating linguistic knowledge acquisition of participants and are commonly used in previous methods [79, 85].

We evaluate the performance of our methods on three multiple-choice **science** QA datasets: ARC-Easy, ARC-Challenge, and OpenBookQA. ARC-Challenge and ARC-easy originate from the same set of exam problems collected from multiple sources. ARC-Challenge contains questions answered incorrectly by both a retrieval-based method and a word co-occurrence method, and the remaining questions form ARC-Easy. Questions in OpenBookQA are

Dataset	Train	Dev	Test	Total
RACE	87,866	4,887	4,934	97,687
ARC-Easy	2,251	570	2,376	5,197
ARC-Challenge	1,119	299	1,172	2,590
OpenBookQA	4,957	500	500	5,957

Table 5.3: The number of questions in RACE and the multiple-choice subject-area QA datasets for evaluation: ARC-Easy, ARC-Challenge, and OpenBookQA.

Dataset	Dev	Test
RACE-M	76.7	76.6
RACE-H	71.0	70.1
RACE-M + RACE-H	72.7	72.0

Table 5.4: Accuracy (%) of the pre-fine-tuned model on the RACE dataset, which contains two subsets: RACE-M and RACE-H, representing problems collected from **m**iddle and **h**igh school language exams, respectively.

crowdsourced by turkers and then carefully filtered and modified by experts. See the statistics of these datasets in Table 5.3. Note that for OpenBookQA, we do not utilize the accompanying auxiliary reference knowledge bases to ensure a fair comparison with previous work.

5.3.2 Experimental Settings

For the two-step fine-tuning framework, we use the uncased $\text{BERT}_{\text{LARGE}}$ released by [75] as the pre-trained language model. We set the batch size to 24, learning rate to 2×10^{-5}, and the maximal sequence length to 512. When the input sequence length exceeds 512, we truncate the longest sequence among q, o_i, and d_i (defined in Section 5.2.1). We first fine-tune $\text{BERT}_{\text{LARGE}}$ for five epochs on RACE to get the pre-fine-tuned model and then further fine-tune the model for eight epochs on the target QA datasets in scientific domains. We show the accuracy of the pre-fine-tuned model on RACE in Table 5.4.

We use the noun phrase chunker in spaCy[2] to extract entity mentions. For information retrieval, we use the version 7.4.0 of Lucene [78] and set the maximum number of the retrieved sentences K to 50. We use the stop word list from NLTK [86].

In addition, we design two slightly different settings for information retrieval. In **setting 1**, the original reference corpus of each dataset is independent. Formally, for each dataset

[2] `https://spacy.io/`.

$x \in D$, we perform information retrieval based on the corresponding original reference corpus of x and/or the external corpus generated based on problems in x, where $D = \{\text{ARC-Easy}, \text{ARC-Challenge}, \text{OpenBookQA}\}$. In **setting 2**, all original reference corpora are integrated to further leverage external in-domain knowledge. Formally, for each dataset $x \in D$, we conduct information retrieval based on the given reference corpus of D and/or the external corpus generated based on problems in D instead of x.[3].

5.3.3 Baselines

Here we only briefly introduce three baselines (i.e., GPT^2, RS^2, and BERT^2) that all fine-tune a pre-trained language model on downstream tasks without substantial modifications to model architectures, which achieve remarkable success on many question answering tasks. Following the two-step fine-tuning framework (Section 5.2.1), **all** three strong baselines use RACE in the first fine-tuning stage for a fair comparison. We will discuss the impact of pre-fine-tuning on baseline model performance in Section 5.3.8, noting that pre-fine-tuning is not the contribution of this work.

GPT^2: This baseline is based on fine-tuning a generative pre-trained transformer (GPT) language model [74] instead of BERT [75].

RS^2: Based on GPT, general reading strategies (RS) [79] are applied during the fine-tuning stage such as adding a trainable embedding into the text embedding of tokens relevant to the question and candidate answer options.

BERT^2: Based on BERT, this baseline is an exact implementation described in Section 5.2.1.

5.3.4 Main Results

We see consistent improvements in accuracy across all tasks after we enrich the reference corpus with relevant texts from Wikipedia to form new reference documents (i.e., RC + EC and IRC + IEC in Table 5.5). Moreover, using only the extracted external corpus to perform information retrieval for reference document generation can achieve reasonable performance compared to using the original reference corpus, especially on the OpenBookQA dataset (62.2% vs. 64.8% under setting 1 and 63.0% vs. 65.0% under setting 2). This indicates that we can extract reliable and relevant texts from external open-domain resources such as Wikipedia via linked entities mentioned in Section 5.2.2. Moreover, using the integrated corpus (i.e., setting 2) consistently boosts the performance. Since the performance in setting

[3] `https://github.com/nlpdata/external`.

Method	ARC-E	ARC-C	OBQA
IR [69]	62.6	20.3	–
Odd-One-Out [70]	–	–	50.2
DGEM [87]	59.0	27.1	24.4
KG^2 [72]	–	31.7	–
AIR [88]	58.4	26.6	–
NCRF++ [84]	52.2	33.2	–
TriAN++ [73]	–	33.4	–
Two Stage Inference [89]	61.1	26.9	–
ET-RR [76]	–	36.6	–
GPT^2 [74, 79]	57.0	38.2	52.0
RS^2 [79]	66.6	40.7	55.2
Our BERT-Based Implementations			
Setting 1			
Reference Corpus (RC) (i.e., BERT^2)	71.9	44.1	64.8
External Corpus (EC)	65.0	39.4	62.2
RC + EC	73.3	45.0	65.2
Setting 2			
Integrated Reference Corpus (IRC)	73.2	44.8	65.0
Integrated External Corpus (IEC)	68.9	40.1	63.0
IRC + IEC	**74.7**	46.1	67.0
IRC + MD	69.4	50.7	67.4
IRC + IEC + MD	72.3	**53.7**	**68.0**
Human Performance	–	–	91.7

Table 5.5: Accuracy (%) on the test sets of ARC-Easy, ARC-Challenge, and OpenBookQA datasets. RACE is used in the pre-fine-tuning stage for all the tasks (Section 5.2.1). MD stands for fine-tuning on **m**ultiple target **d**atasets simultaneously (Section 5.2.3). All results are single-model performance. GPT^2, RS^2, and BERT^2 are baselines that use two-step fine-tuning (Section 5.3.3). ARC-E: ARC-Easy; ARC-C: ARC-Challenge; OBQA: OpenBookQA.

2 (integrated corpus) is better than that in setting 1 (independent corpus) based on our experiments, we take **setting 2** by default for discussions unless explicitly specified.

We see further improvements on ARC-Challenge and OpenBookQA, by fine-tuning the pre-fine-tuned model on multiple target datasets (i.e., ARC-Easy, ARC-Challenge, and OpenBookQA). However, we do not see a similar gain on ARC-Easy by increasing the number of in-domain training instances. We will further discuss it in Section 5.3.6.

Our best models (i.e., IRC + IEC for ARC-Easy and IRC + IEC + MD for ARC-Challenge and OpenBookQA) outperform the strong baseline BERT^2 introduced in Section 5.2.1 (74.7% vs. 71.9% on ARC-Easy, 53.7% vs. 44.1% on ARC-Challenge, and 68.0% vs. 64.8%

Question	Answer Options	Sentence(s) From Wikipedia
What boils at the boiling point?	A. ***Kool-Aid***. ✓ B. Cotton. C. Paper Towel. D. Hair.	***Kool-Aid*** is known as Nebraska's official soft drink. Common types of drinks include plain drinking ***water***, milk, coffee, tea, hot chocolate, juice and ***soft drinks***.
Forest fires occur in many areas due to ***drought conditions***. If the drought conditions continue for a long period of time, which might cause the repopulation of trees to be threatened?	A. a decrease in the ***thickness of soil***. ✓ B. a decrease in the amount of erosion. C. an increase in the bacterium population. D. an increase in the production of oxygen and fire.	It is highly resistant to ***drought conditions***, and provides excellent fodder; and has also been used in controlling ***soil erosion***, and as revegetator, often after ***forest fires***.
Juan and LaKeisha roll a few objects down a ramp. They want to see which object rolls the farthest. What should they do so they can repeat their ***investigation?***	A. Put the objects in groups. B. Change the height of the ramp. C. Choose different objects to roll. D. ***Record*** the details of the ***investigation***. ✓	The use of measurement developed to allow ***recording*** and comparison of ***observations*** made at different times and places, by different people.
Which statement best explains why the sun appears to ***move across the sky*** each day?	A. The sun revolves around Earth. B. Earth rotates around the sun. C. The sun revolves on its axis. D. ***Earth rotates*** on its ***axis***. ✓	***Earth's rotation*** about its ***axis*** causes the fixed stars to apparently ***move across the sky*** in a way that depends on the observer's latitude.

Table 5.6: Examples of corrected errors using the reference corpus enriched by the sentences from Wikipedia.

on OpenBookQA), which already beats the previous state-of-the-art model RS^2. In the remaining sections, we analyze our models and discuss the impacts of external knowledge from various aspects.

5.3.5 Impact of External Knowledge from an Open-Domain Resource

Table 5.6 shows some examples of errors produced by IRC (Table 5.5) that do not leverage external knowledge from open-domain resources. These errors can be corrected by enriching the reference corpus with external sentences extracted from Wikipedia (IRC + IEC in Table 5.5). In the first example, the correct answer option *"Kool-Aid"* never appears in the original reference corpus. As a result, without external background knowledge, it is less likely to infer that *"Kool-Aid"* refers to liquid (can boil) here.

In addition to performing information retrieval on the enriched reference *corpus*, we investigate an alternative approach that uses entity identification and linking to directly enrich the reference *document* for each (question, answer option) pair. More specifically, we apply

Task	Wiki	OBQA	ARC	Total
ARC-E	20.8	0.4	78.7	1,039,059
ARC-C	21.5	0.4	78.2	517,846
OBQA	20.6	1.1	78.3	1,191,347

Table 5.7: Percentage (%) of retrieved sentences from each source. Wiki: Wikipedia; Total: total number of retrieved sentences for all (question, answer option) pairs in a single task. ARC-Easy and ARC-Challenge share the same original reference corpus.

First 4	Last 4	Accuracy	# Epochs
ARC-C	ARC-E	69.4	8
OBQA	ARC-E	70.9	8
ARC-C + OBQA	ARC-E	72.6	8
ARC-E	-	72.9	4
ARC-E	ARC-E	74.7	8

Table 5.8: Accuracy (%) on the ARC-Easy test set. The first four epochs are fine-tuned using the dataset(s) in the first column. The last four epochs are fine-tuned using the dataset in the second column. # Epochs: the total number of epochs.

entity extraction and linking to each (question, answer option) pair (q, o_i) and extract sentences from Wikipedia based on the linked entities. These extracted sentences are appended to the reference documents d_i of (q, o_i) directly. We still keep up to K (i.e., 50) sentences per document. We observe that this direct appending approach generally cannot outperform the reference corpus enrichment approach described in Section 5.2.2.

We report the statistics of the sentences (without redundancy removal) extracted from each source in Table 5.7, used as inputs to our methods IRC + IEC and IRC + IEC + MD in Table 5.5. As the original reference corpus of OpenBookQA is made up of 1,326 sentences, fewer retrieved sentences are extracted from its reference corpus for all tasks compared to other sources.

5.3.6 Impact of External Knowledge from In-Domain Data

Compared to fine-tuning the pre-fine-tuned model on a single multiple-choice subject-area QA dataset, we observe improvements in accuracy by fine-tuning on multiple in-domain datasets (MD) simultaneously (Section 5.2.3) for ARC-Challenge and OpenBookQA. In particular, we see a dramatic gain on the ARC-Challenge dataset (from 46.1% to 53.7%) as shown in Table 5.5.

However, MD leads to a performance drop on ARC-Easy. We hypothesize that other commonly adopted approaches may also lead to performance drops. To verify that, we explore another way of utilizing external knowledge for ARC-Easy by first fine-tuning the pre-fine-tuned model for four epochs on external in-domain data (i.e., ARC-Challenge, OpenBookQA, or ARC-Challenge + OpenBookQA) and then further fine-tuning for four epochs on ARC-Easy. As shown in Table 5.8, we also observe that compared to only fine-tuning on ARC-Easy, fine-turning on external in-domain data hurts the performance. The consistent performance drops across the two methods of using MD on ARC-Easy are perhaps due to an intrinsic property of the tasks themselves – the question-answer instances in ARC-Easy are relatively simpler than those in ARC-Challenge and OpenBookQA. Introducing relatively complex problems from ARC-Challenge and OpenBookQA may hurt the final performance on ARC-Easy. As mentioned earlier, compared to questions in ARC-Easy, questions in ARC-Challenge are less likely to be answered correctly by retrieval-based or word co-occurrence methods. We argue that questions in the ARC-Challenge tend to require more external knowledge for reasoning, similar to the observation of [90] (30.0% vs. 20.0%).

5.3.7 Discussions about Question Types and Remaining Challenges

We use the human annotations such as required reasoning skills (i.e., *word matching*, *paraphrasing*, *knowledge*, *meta/whole*, and *math/whole*) and validity of questions in ARC-Easy and ARC-Challenge released by [90] to analyze the impacts of external knowledge on instances in various categories. Sixty instances are annotated for each dataset. We refer readers to [90] for detailed definitions of each category. We do not report the accuracy for *math/whole* as no annotated question in ARC belongs to this category.

We compare the BERT^2 baseline in Table 5.5 that only uses the original reference corpus of a given end task with our best model. As shown in Table 5.9, by leveraging external knowledge from in-domain datasets (instances and reference corpora) and open-domain texts, we observe consistent improvements on most of the categories. Based on these experimental results on the annotated subset, we may assume it could be a promising direction to further improve challenging multiple-choice subject-area QA tasks through exploiting high-quality external knowledge besides designing task-specific models for different types of questions [68].

We also analyze the instances that our approach fails to answer correctly in the OpenBookQA development set to study the remaining challenges. It might be promising to identify the relations among entities within an answer option. For example, our current model mistakenly selects the answer option *"the sun orbits the earth"* associated with the question *"Revolution happens when ?"* probably because *"sun"*, *"orbits"*, and *"earth"* frequently co-

Question Type	ARC-E		ARC-C	
	BERT²	Ours	BERT²	Ours
Word Matching	81.3	**85.4**	30.4	**73.9**
Paraphrasing	90.9	90.9	46.7	**66.7**
Knowledge	58.3	**83.3**	44.4	**55.6**
Math/Logic	100.0	100.0	33.3	33.3
Valid	80.0	**86.0**	36.1	**66.7**
Invalid	50.0	**80.0**	41.7	41.7
Easy	80.0	**90.0**	33.3	**53.3**
Hard	70.0	**80.0**	43.3	**60.0**

Table 5.9: Accuracy (%) by different categories on the annotated test sets of ARC-Easy and ARC-Challenge, which are released by sugawara2018makes.

occur in our generated reference document, though these entities such as *"revolution"* are successfully linked to their corresponding Wikipedia pages in the astronomy field.

Besides, we might also need to identify causal relations between events. For example, given the question *"The type of climate change known as anthropogenic is caused by this"*, our model mistakenly predicts another answer option *"forest fires"* with its associated contexts *"climate change has caused the island to suffer more frequent severe droughts, leading to large forest fires"*, instead of the real cause *"humanity"* supported by *"the problem now is with anthropogenic climate change—that is, climate change caused by human activity, which is making the climate change a lot faster than it normally would"*.

5.3.8 Discussions about Pre-Fine-Tuning

Previous work [75] has shown that fine-tuning $\text{BERT}_{\text{LARGE}}$ on small datasets can be sometimes unstable. Additionally, [79] show that fine-tuning GPT [74] that is pre-fine-tuned on RACE can dramatically improve the performance of relatively small multiple-choice tasks. Here we only use the BERT² baseline for a brief discussion. We have a similar observation: we can obtain more stable performance on the target datasets by first fine-tuning BERT on RACE (language exams), and we see consistent performance improvements on all the evaluated science QA datasets. As shown in Figure 5.2, we see that the performance drops dramatically without using pre-fine-tuning on the RACE dataset.

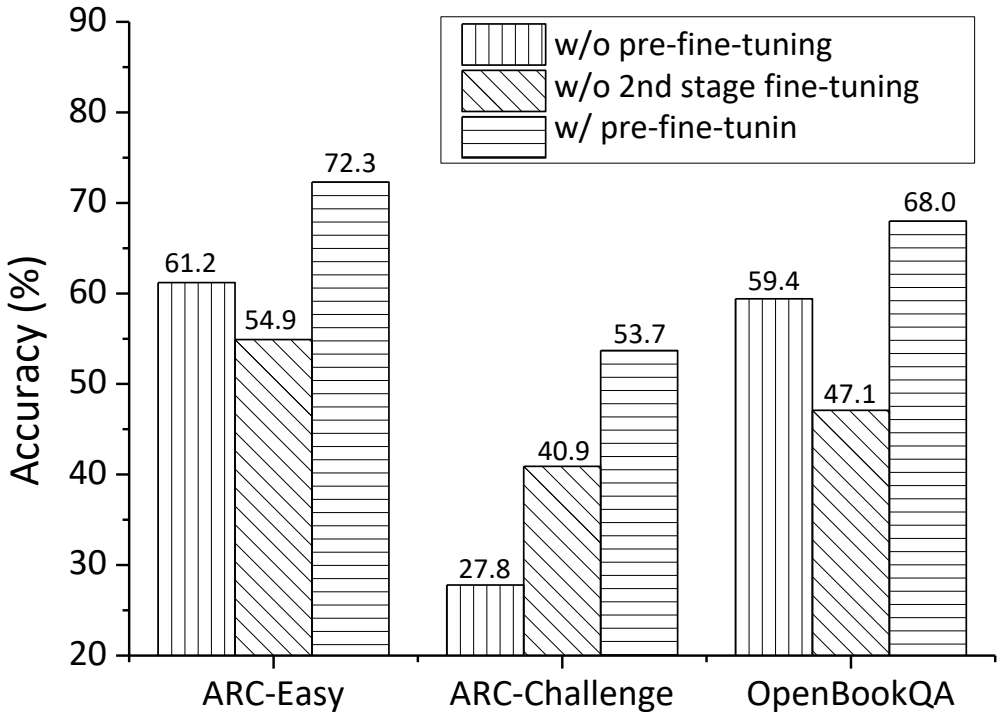


Figure 5.2: Accuracy (%) on the test sets of evaluation tasks with and without the pre-fine-tuning stage (2nd stage fine-tuning: fine-tune the pre-fine-tuned model on target science question answering datasets).

5.4 SUMMARY

In this chapter, we focus on how to incorporate external knowledge into a pre-trained model to improve subject-area QA tasks that require background knowledge. We exploit two sources of external knowledge through: enriching the original reference corpus with relevant texts from open-domain Wikipedia and using additional in-domain QA datasets (instances and reference corpora) for training. Experimental results on ARC-Easy, ARC-Challenge, and OpenBookQA show the effectiveness of our simple method. The promising results also demonstrate the importance of unstructured external knowledge for subject-area QA.

CHAPTER 6: RELATED WORK

6.1 AMR BASED NATURAL LANGUAGE PROCESSING

In this thesis, we demonstrate that AMR is an appropriate and elegant way to acquire, select, represent and organize deeper knowledge in text. Together with our novel utilization of the rich structures in merged KBs, the whole framework carries rich enough evidence for effective entity linking, without the need for any labeled data, collective inference, or sophisticated similarity computation methods. AMR has been applied for many other NLP tasks: Text Summarization [91], Combinatory Categorial Grammar parsing [92, 93], Event Detection [94, 95] and Language Generation [96, 97].

6.2 CROSS-LINGUAL ENTITY LINKING

NIST TAC-KBP Tri-lingual entity linking [98] focused on three languages: English, Chinese and Spanish. [44] extended it to developed a cross-lingual entity linking system for 21 languages. But their methods required labeled data and name transliteration. We share the same goal as [99] to extend cross-lingual entity linking to all languages in Wikipedia. They exploited Wikipedia links to train a supervised linker. We mine reliable word translations from cross-lingual Wikipedia titles, which enables us to adopt unsupervised English entity linking techniques such as [16] to directly link translated English name mentions to English KB. [100] built a massively multilingual corpus using resources including Wikipedia. To the best of our knowledge, our work covers the largest number of languages for both name tagging and linking. Recent deep neural networks based methods for cross-lingual entity linking [101, 102] rely on a large amount of manually annotated data.

6.3 COLLECTIVE ENTITY LINKING

In most recent collective inference methods for Entity Linking (e.g., [7, 14, 30, 37, 39, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112]), the target entity mention's "collaborators" may simply include all mentions which co-occur in the same discourse (sentence, paragraph or document) [37, 113]. But this approach usually introduces many irrelevant mentions, and it's very difficult to automatically determine the scope of discourse. In contrast, some recent work exploited more restricted measures by only choosing those mentions which are topically related [15, 114], bear a relation from a fixed set [24], coreferential [113, 115], socially

related [15, 115], dependent [116], or a combination of these through meta-paths [115]. These measures can collect more precise collaborators but suffer from low coverage of pre-defined information templates and the unsatisfying quality of state-of-the-art coreference resolution, relation and event extraction. We demonstrate that AMR is an appropriate and elegant way to acquire, select, represent and organize deeper knowledge in text. Together with our novel utilization of the rich structures in merged KBs, the whole framework carries rich enough evidence for effective EL, without the need for any labeled data, collective inference, or sophisticated similarity.

6.4 CROSS-LINGUAL EMBEDDING LEARNING

[31] first observes that word embedding spaces have similar geometric arrangements across languages. They propose to use this property to learn a linear mapping between two spaces. After that, several methods attempt to improve the mapping quality [33, 117, 118, 119, 120, 121]. Recent methods have shown that it is possible to derive cross-lingual word embedding from unaligned corpora in an unsupervised framework [34, 122, 123].

Another strategy for cross-lingual word embedding learning is to combine monolingual and cross-lingual training objectives in a joint fashion [124, 125, 126, 127, 128]. Compared to our direct mapping approach, these methods generally require a large size of parallel data.

Previous work on cross-lingual joint entity and word embedding methods largely neglects unlinkable entities [29] and heavily relies on parallel or comparable sentences [129]. [29] applies a similar approach to generate code-switched data from Wikipedia, but their framework does not reserve entities in the source language. Using all aligned entities as a dictionary, they adopt canonical correlation analysis to project two embedding spaces into one. In contrast, we only choose salient entities as anchors to learn a linear mapping. [129] generates comparable data via distant supervision over multilingual knowledge bases, and uses two types of regularizers, entity regularizer, and sentence regularizer, to align cross-lingual words and entities. Further, they design knowledge attention and cross-lingual attention to refine the alignment. Essentially, they train cross-lingual embedding jointly, while we align two embedding that trained independently. Moreover, compared to their approach that relies on comparable data, aligned entities are easier to acquire.

6.5 WIKIPEDIA MARKUP BASED SILVER STANDARD GENERATION

Our data generation method for embedding learning and silver-standard name tagging annotation is mainly inspired from previous work that leverages Wikipedia markups to train name taggers [56, 130, 131, 132, 133, 134, 135]. Most of these previous approaches require manual annotations to assign types to a certain amount of Wikipedia entries as seeds in order to train the tagger. In contrast, we exploit AMR corpus to train an English typing system and then transfer the labels from English Wikipedia entries to all other languages. Most of these previous methods manually classify many English Wikipedia entries into pre-defined entity types. In contrast, our approach doesn't need any manual annotations or language-specific features, while generates both coarse-grained and fine-grained types. Moreover, previous work on silver-standard annotation generation only focuses on name tagging, while we also include cross-lingual entity linking into the framework and extend it to all languages in Wikipedia.

Some previous work including [136, 137, 138] exploit semi-supervised methods to save annotation cost. We observe that self-training can provide further gains when the training data contains a certain amount of noise.

Many fine-grained entity typing approaches [139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149] also create annotations based on Wikipedia anchor links. Compared to these methods, our framework performs both name identification and typing, and takes advantage of richer structures in the KBs. One unique challenge to develop name taggers for many morphological-rich languages is to perform effective morphological analysis. Previous work on Arabic name tagging [135] extract entity titles as gazetteers for stemming, and thus it cannot handle unknown names. We propose a new method to derive generalizable affixes for morphologically rich language based on Wikipedia markups.

6.6 WIKIPEDIA AS BACKGROUND FEATURES FOR IE

Wikipedia pages have been used as additional features to improve various Information Extraction (IE) tasks, including name tagging [150], coreference resolution [151], relation extraction [152] and event extraction [153]. Other automatic name annotation generation methods have been proposed, including KB driven distant supervision [154, 155, 156] and cross-lingual projection [157, 158, 159, 160, 161, 162].

6.7 MULTILINGUAL NAME TAGGING

Some recent research [62, 63, 163, 164, 165] under the DARPA LORELEI program focuses on developing (almost) language universal resources and techniques for low-resource languages. These approaches require English annotations for projection [63], some input from a native speaker for each language, either through manual annotations [163], or a linguistic survey to acquire language-specific rules and patterns [62], or comprehensive "Chinese Room " style annotation interfaces [166] for non-speakers to perform name annotation [167, 168]. We take full advantage of Wikipedia markups (typing and morphology analysis), cross-lingual links, and DBPedia properties. Without using any manual annotations, our name taggers outperform previous methods on the same data sets for many languages.

6.8 SUBJECT-AREA QA TASKS AND METHODS

As there is not a clear distinction between QA and machine reading comprehension (MRC) tasks, for convenience we call a task in which there is no reference document provided for each instance as a QA task. In this thesis, we focus on multiple-choice subject-area QA tasks, where the in-domain reference corpus does not provide sufficient relevant content on its own to answer a significant portion of the questions [68, 69, 70, 169, 170]. In contrast to other types of QA scenarios [171, 172, 173, 174, 175], in this setting: (1) the reference corpus does not reliably contain text spans from which the answers can be drawn, and (2) it does not provide sufficient information on its own to answer a significant portion of the questions. Thus they are suitable for us to study how to exploit external knowledge for QA.

Our work follows the general framework of discriminatively fine-tuning a pre-trained language model such as GPT [74] and BERT [75] on QA tasks [74, 75, 176, 177]. As shown in Table 5.5, the baseline based on BERT already outperforms previous state-of-the-art methods designed for subject-area QA tasks [76, 79, 88, 89].

6.8.1 Utilization of External Knowledge for Subject-Area QA

Previous studies have explored many ways to leverage structured knowledge to solve questions in subject areas such as science exams. Many researchers investigate how to directly or indirectly use automatically constructed knowledge bases/graphs from reference corpora [71, 72, 178, 179] or existing external general knowledge graphs [73, 84, 85, 180, 181, 182] such as ConceptNet [183]. However, for subject-area QA, unstructured knowledge is seldom considered in previous studies, and it is still not clear the usefulness of this kind of knowledge.

As far as we know, for subject-area QA tasks, this is the first attempt to impart sources of external unstructured knowledge into one state-of-the-art pre-trained language model, and we are among the first to investigate the effectiveness of the external unstructured texts in Wikipedia [89] and additional in-domain QA data.

6.8.2 Utilization of External Knowledge for Other Types of QA and MRC

For both QA and MRC tasks in which the majority of answers are extractive such as SQuAD [184] and TriviaQA [173], previous work has shown that it is useful to introduce external open-domain QA instances and textual information from Wikipedia by first retrieving relevant documents in Wikipedia and then running a MRC model to extract a text span from the documents based on the question [185, 186, 187, 188, 189].

Based on Wikipedia, we apply concept identification and linking to enrich QA reference corpora, which has not been explored before. Compared to previous data argumentation studies for other types of QA tasks [190], differences exist in: (1) we focus on in-domain data and discuss the impacts of the difficulties of additional in-domain instances on a target task; (2) we are the first to show it is useful to merge reference corpora from different in-domain subject-area QA tasks.

CHAPTER 7: CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this thesis, we have introduced two semantic representations for entities, symbolic semantics based (Chapter 2) and distributed semantics based (Chapter 3), to perform cross-lingual entity extraction and linking. We also propose a novel framework to combine them for fine-grained entity extraction and cross-lingual linking. Our framework has significantly extended the entity extraction and linking capabilities from seven coarse-grained types to thousands of fine-grained types, and from several high-resource languages to 300 Wikipedia languages. Despite of these successes, many unique challenges remain. Since the silver-standard training data is mainly derived from Wikipedia, the performance of entity extraction heavily relies on the amount of available Wikipedia entries for a certain language. Our current method achieves up to 76% F-score for entity extraction for non-Wikipedia data, which is much lower than English. To further improve the performance we would need to prepare training data with higher quality and incorporate more language-specific features. In addition, our data sets contain many popular entities in news and social media. When applying our techniques to real-world domains such as conversational systems, we expect to face new challenges due to a massive amount of new entities emerging every day. To tackle this challenge we need to be able to rapidly construct profiles of new entities, and construct a personal profile of the user who asks about the entity and perform personalized entity linking.

CHAPTER 8: REFERENCES

[1] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation for sembanking," in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/W13-2322 pp. 178–186.

[2] F. Mahdisoltani, J. Biega, and F. M. Suchanek, "Yago3: A knowledge base from multilingual wikipedias," in *Proceedings of the Conference on Innovative Data Systems Research*, 2015.

[3] H. Ji, J. Nothman, B. Hachey, and R. Florian, "Overview of tac-kbp2015 tri-lingual entity discovery and linking," in *Proc. Text Analysis Conference (TAC2015)*, 2015.

[4] D. Milne and I. H. Witten, "Learning to link with wikipedia," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1458082.1458150 pp. 509–518.

[5] Z. Syed, T. Finin, and A. Joshi, "Wikipedia as an ontology for describing documents," in *Proc. International Conference on Weblogs and Social Media (ICWSM 2008)*, 2008.

[6] H. Srinivasan, J. Chen, and R. Srihari, "Cross document person name disambiguation using entity profiles," in *Proc. Text Analysis Conference (TAC 2009)*, 2009.

[7] Z. Kozareva, K. Voevodski, and S. Teng, "Class label enhancement via related instances," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011. [Online]. Available: http://aclweb.org/anthology/D11-1011 pp. 118–128.

[8] W. Zhang, J. Su, and C.-L. Tan, "A wikipedia-lda model for entity linking with batch size changing instance selection," in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 2011. [Online]. Available: http://aclweb.org/anthology/I11-1063 pp. 562–570.

[9] I. Anastácio, B. Martins, and P. Calado, "Supervised learning for linking named entities to knowledge base entries," in *Proc. Text Analysis Conference (TAC 2011)*, 2011.

[10] T. Cassidy, Z. Chen, J. Artiles, H. Ji, H. Deng, L. Ratinov, J. Zheng, J. Han, and D. Roth, "Cuny-uiuc-sri tac-kbp2011 entity linking system description," in *Proc. Text Analysis Conference (TAC 2011)*, 2011.

[11] G. Pink, W. Radford, W. Cannings, A. Naoum, J. Nothman, D. Tse, and J. Curran, "Sydney cmcrc at tac 2013," in *Proc. Text Analysis Conference (TAC 2013)*, 2013.

[12] S. Gao, Y. Cai, S. Li, Z. Zhang, Y. Guan, J.and Li, H. Zhang, W. Xu, and J. Guo, "Pris at tac2010 kbp track," in *Proc. Text Analysis Conference (TAC 2010)*, 2010.

[13] Z. Chen and H. Ji, "Collaborative ranking: A case study on entity linking," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011. [Online]. Available: http://aclweb.org/anthology/D11-1071 pp. 771–781.

[14] Z. Chen and H. Ji, "Collaborative ranking: A case study on entity linking," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011. [Online]. Available: http://aclweb.org/anthology/D11-1071 pp. 771–781.

[15] T. Cassidy, H. Ji, L.-A. Ratinov, A. Zubiaga, and H. Huang, "Analysis and enhancement of wikification for microblogs with context expansion," in *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, 2012. [Online]. Available: http://aclweb.org/anthology/C12-1028 pp. 441–456.

[16] X. Pan, T. Cassidy, U. Hermjakob, H. Ji, and K. Knight, "Unsupervised entity linking with abstract meaning representation," in *Proceedings of the NAACL-HLT*, Denver, CO, 2015. [Online]. Available: http://www.aclweb.org/anthology/N15-1119

[17] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, "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)*. Association for Computational Linguistics, 2017. [Online]. Available: http://aclweb.org/anthology/P17-1178 pp. 1946–1958.

[18] X. Pan, T. Gowda, H. Ji, J. May, and S. Miller, "Cross-lingual joint entity and word embedding to improve entity linking and machine translation," in *Submitted to ACL 2019*, 2019.

[19] S. S. Pradhan and N. Xue, "Ontonotes: The 90% solution," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*. Association for Computational Linguistics, 2009. [Online]. Available: http://aclweb.org/anthology/N09-4006 pp. 11–12.

[20] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational Linguistics*, vol. 31, no. 1, 2005. [Online]. Available: http://aclweb.org/anthology/J05-1004

[21] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman, "The nombank project: An interim report," in *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, 2004. [Online]. Available: http://aclweb.org/anthology/W04-2705

[22] H. Ji, R. Grishman, Z. Chen, and P. Gupta, "Cross-document event extraction and tracking: Task, evaluation, techniques and challenges," in *Proceedings of the International Conference RANLP-2009*. Association for Computational Linguistics, 2009. [Online]. Available: http://aclweb.org/anthology/R09-1032 pp. 166–172.

[23] O. Medelyan and C. Legg, "Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense," in *Proceedings of the AAAI 2008 Workshop on WIKIAI*, Chicago, IL, 2008. [Online]. Available: http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-003.pdf

[24] X. Cheng and D. Roth, "Relational inference for wikification," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/D13-1184 pp. 1787–1796.

[25] J. Flanigan, S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith, "A discriminative graph-based parser for the abstract meaning representation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/P14-1134 pp. 1426–1436.

[26] Z. S. Harris, "Distributional structure," *WORD*, vol. 10, no. 2-3, pp. 146–162, 1954. [Online]. Available: https://doi.org/10.1080/00437956.1954.11659520

[27] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, "Problems with evaluation of word embeddings using word similarity tasks," in *Proc. of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 2016.

[28] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph and text jointly embedding," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. [Online]. Available: http://www.aclweb.org/anthology/D14-1167 pp. 1591–1601.

[29] C.-T. Tsai and D. Roth, "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*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/N16-1072 pp. 589–598.

[30] I. Yamada, H. Shindo, H. Takeda, and Y. Takefuji, "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*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/K16-1025 pp. 250–259.

[31] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation." *CoRR*, 2013.

[32] C. Xing, D. Wang, C. Liu, and Y. Lin, "Normalized word embedding and orthogonal transform for bilingual word translation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/N15-1104 pp. 1006–1011.

[33] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *CoRR*, vol. abs/1702.03859, 2017. [Online]. Available: http://arxiv.org/abs/1702.03859

[34] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[35] S. Cucerzan, "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)*, 2007. [Online]. Available: http://aclweb.org/anthology/D07-1074

[36] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011. [Online]. Available: http://aclweb.org/anthology/D11-1072 pp. 782–792.

[37] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to wikipedia," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011. [Online]. Available: http://aclweb.org/anthology/P11-1138 pp. 1375–1384.

[38] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani, "Learning relatedness measures for entity linking," in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, ser. CIKM '13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505711 pp. 139–148.

[39] X. Ling, S. Singh, and D. S. Weld, "Design challenges for entity linking," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 315–328, 2015. [Online]. Available: http://aclweb.org/anthology/Q15-1023

[40] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: http://arxiv.org/abs/1310.4546

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[42] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[43] J. Mayfield, D. Lawrie, P. McNamee, and D. W. Oard, "Building a cross-language entity linking collection in twenty-one languages," in *Multilingual and Multimodal Information Access Evaluation: Second International Conference of the Cross-Language Evaluation Forum*, 2011.

[44] P. McNamee, J. Mayfield, D. Lawrie, D. Oard, and D. Doermann, "Cross-language entity linking," in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 2011. [Online]. Available: http://aclweb.org/anthology/I11-1029 pp. 255–263.

[45] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations*, ser. ICLR '17, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[46] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[47] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1247–1256.

[48] C. Wang, S. Pradhan, X. Pan, H. Ji, and N. Xue, "CAMR at SemEval-2016 task 8: An extended transition-based AMR parser," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016. [Online]. Available: https://www.aclweb.org/anthology/S16-1181 pp. 1173–1178.

[49] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1376616.1376746 pp. 1247–1250.

[50] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin, "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking," in *Proceeddings ACL2008*, 2008.

[51] A. Mahmoudi, M. Arabsorkhi, and H. Faili, "Supervised morphology generation using parallel corpus," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. INCOMA Ltd. Shoumen, BULGARIA, 2013. [Online]. Available: http://aclweb.org/anthology/R13-1053 pp. 408–414.

[52] M. Ahlberg, M. Forsberg, and M. Hulden, "Paradigm classification in supervised learning of morphology," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/N15-1107 pp. 1024–1029.

[53] S.-A. Grönroos, S. Virpioja, P. Smit, and M. Kurimo, "Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/C14-1111 pp. 1177–1185.

[54] T. Ruokolainen, O. Kohonen, K. Sirts, S.-A. Grönroos, M. Kurimo, and S. Virpioja, "A comparative study of minimally supervised morphological segmentation," *Computational Linguistics*, vol. 42, no. 1, pp. 91–120, 2016. [Online]. Available: http://aclweb.org/anthology/J16-1003

[55] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," in *27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, June 1989. [Online]. Available: https://www.aclweb.org/anthology/P89-1010 pp. 76–83.

[56] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from wikipedia," *Artificial Intelligence*, vol. 194, pp. 151–175, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2012.03.006

[57] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/N16-1030 pp. 260–270.

[58] J. Guo, W. Che, H. Wang, and T. Liu, "Revisiting embedding features for simple semi-supervised learning," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/D14-1012 pp. 110–120.

[59] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proceedings of the Biennial GSCL Conference 2009*, 2009.

[60] H. Li, H. Ji, H. Deng, and J. Han, "Exploiting background information networks to enhance bilingual event extraction through topic modeling," in *Proceedings of International Conference on Advances in Information Mining and Management (IMMM2011)*, 2011.

[61] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/P14-5010 pp. 55–60.

[62] B. Zhang, X. Pan, T. Wang, A. Vaswani, H. Ji, K. Knight, and D. Marcu, "Name tagging for low-resource incident languages based on expectation-driven learning," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/N16-1029 pp. 249–259.

[63] C.-T. Tsai, S. Mayhew, and D. Roth, "Cross-lingual named entity recognition via wikification," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/K16-1022 pp. 219–228.

[64] Y. Lin and H. Ji, "An attentive fine-grained entity typing model with latent type representation," in *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)*, 2019.

[65] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

[66] D. S. McNamara, I. B. Levinstein, and C. Boonthum, "iSTART: Interactive strategy training for active reading and thinking," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, 2004. [Online]. Available: https://link.springer.com/article/10.3758/BF03195567

[67] L. Salmerón, W. Kintsch, and J. J. Caãs, "Reading strategies and prior knowledge in learning from hypertext," *Memory & Cognition*, vol. 34, no. 5, 2006. [Online]. Available: https://link.springer.com/article/10.3758/BF03193262

[68] P. Clark, O. Etzioni, T. Khot, A. Sabharwal, O. Tafjord, P. D. Turney, and D. Khashabi, "Combining retrieval, statistics, and inference to answer elementary science questions." in *Proceedings of the AAAI*, Phoenix, AZ, 2016. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/11963/11990

[69] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? Try ARC, the AI2 reasoning challenge," *arXiv preprint*, vol. cs.CL/1803.05457v1, 2018. [Online]. Available: https://arxiv.org/abs/1803.05457v1

[70] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *Proceedings of the EMNLP*, Brussels, Belgium, 2018. [Online]. Available: http://aclweb.org/anthology/D18-1260

[71] T. Khot, A. Sabharwal, and P. Clark, "Answering complex questions using open information extraction," in *Proceedings of the ACL*, Vancouver, Canada, 2017. [Online]. Available: http://www.aclweb.org/anthology/P17-2049

[72] Y. Zhang, H. Dai, K. Toraman, and L. Song, "KGˆ 2: Learning to reason science exam questions with contextual knowledge graph embeddings," *arXiv preprint*, vol. cs.LG/1805.12393v1, 2018. [Online]. Available: https://arxiv.org/abs/1805.12393v1

[73] W. Zhong, D. Tang, N. Duan, M. Zhou, J. Wang, and J. Yin, "Improving question answering by commonsense-based pre-training," *arXiv preprint*, vol. cs.CL/1809.03568v1, 2018. [Online]. Available: https://arxiv.org/abs/1809.03568v1

[74] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," in *Preprint*, 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[75] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proeedings of the NAACL-HLT*, Minneapolis, MN, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805v1

[76] J. Ni, C. Zhu, W. Chen, and J. McAuley, "Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering," in *Proceedings of the NAACL-HLT*, Minneapolis, MN, 2019. [Online]. Available: https://arxiv.org/abs/1808.09492v4

[77] V. Yadav, S. Bethard, and M. Surdeanu, "Alignment over heterogeneous embeddings for question answering," in *Proceedings of the NAACL-HLT*, Minneapolis, MN, 2019. [Online]. Available: https://www.aclweb.org/anthology/N19-1274

[78] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action, Second Edition: Covers Apache Lucene 3.0.* Greenwich, CT: Manning Publications Co., 2010. [Online]. Available: https://dl.acm.org/citation.cfm?id=1893016

[79] K. Sun, D. Yu, D. Yu, and C. Cardie, "Improving machine reading comprehension with general reading strategies," in *Proceedings of the NAACL-HLT*, Minneapolis, MN, 2019. [Online]. Available: https://arxiv.org/abs/1810.13441v2

[80] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale reading comprehension dataset from examinations," in *Proceedings of the EMNLP*, Copenhagen, Denmark, 2017. [Online]. Available: http://www.aclweb.org/anthology/D17-1082

[81] P. Kendeou and P. Van Den Broek, "The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts," *Memory & cognition*, vol. 35, no. 7, 2007. [Online]. Available: https://link.springer.com/article/10.3758/BF03193491

[82] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of the ACL (Demonstrations)*, Baltimore, MD, 2014. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[83] D. Khashabi, T. Khot, A. Sabharwal, and D. Roth, "Learning what is essential in questions," in *Proceedings of the CoNLL 2017*, 2017. [Online]. Available: http://www.aclweb.org/anthology/K17-1010

[84] R. Musa, X. Wang, A. Fokoue, N. Mattei, M. Chang, P. Kapanipathi, B. Makni, K. Talamadupula, and M. Witbrock, "Answering science exam questions using query rewriting with background knowledge," *arXiv preprint*, vol. cs.AI/1809.05726v1, 2018. [Online]. Available: https://arxiv.org/abs/1809.05726v1

[85] L. Wang, M. Sun, W. Zhao, K. Shen, and J. Liu, "Yuanfudao at SemEval-2018 Task 11: Three-way attention and relational knowledge for commonsense machine comprehension," in *Proceedings of the SemEval*, New Orleans, LA, 2018. [Online]. Available: http://aclweb.org/anthology/S18-1120

[86] S. Bird and E. Loper, "NLTK: the natural language toolkit," in *Proceedings of the ACL (Demonstrations)*, Barcelona, Spain, 2004. [Online]. Available: http://www.aclweb.org/anthology/P04-3031

[87] T. Khot, A. Sabharwal, and P. Clark, "SciTail: A textual entailment dataset from science question answering," in *Proceedings of the AAAI*, New Orleans, LA, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/17368

[88] V. Yadav, R. Sharp, and M. Surdeanu, "Sanity check: A strong alignment and information retrieval baseline for question answering," in *Proceedings of the ACM SIGIR*, Ann Arbor, MI, 2018. [Online]. Available: https://dl.acm.org/citation.cfm?id=3210142

[89] G.-S. Pirtoaca, T. Rebedea, and S. Ruseti, "Improving retrieval-based question answering with deep inference models," *arXiv preprint*, vol. cs.CL/1812.02971v2, 2019. [Online]. Available: https://arxiv.org/abs/1812.02971v2

[90] S. Sugawara, K. Inui, S. Sekine, and A. Aizawa, "What makes reading comprehension questions easier?" in *Proceedings of the EMNLP*, Brussels, Belgium, 2018. [Online]. Available: http://aclweb.org/anthology/D18-1453

[91] F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith, "Toward abstractive summarization using semantic representations," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/N15-1114 pp. 1077–1086.

[92] Y. Artzi, K. Lee, and L. Zettlemoyer, "Broad-coverage ccg semantic parsing with amr," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/D15-1198 pp. 1699–1710.

[93] D. K. Misra and Y. Artzi, "Neural shift-reduce ccg semantic parsing," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/D16-1183 pp. 1775–1786.

[94] X. Li, T. H. Nguyen, K. Cao, and R. Grishman, "Improving event detection with abstract meaning representation," in *Proceedings of the First Workshop on Computing News Storylines*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/W15-4502 pp. 11–15.

[95] S. Rao, D. Marcu, K. Knight, and H. Daumé III, "Biomedical event extraction using abstract meaning representation," in *BioNLP 2017*. Association for Computational Linguistics, 2017. [Online]. Available: http://aclweb.org/anthology/W17-2315 pp. 126–135.

[96] J. Flanigan, C. Dyer, N. A. Smith, and J. Carbonell, "Generation from abstract meaning representation using tree transducers," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/N16-1087 pp. 731–739.

[97] N. Pourdamghani, K. Knight, and U. Hermjakob, "Generating english from abstract meaning representations," in *Proceedings of the 9th International Natural Language Generation conference*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/W16-6603 pp. 21–25.

[98] H. Ji, J. Nothman, and H. T. Dang, "Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end kbp," in *Proceedings of the Text Analysis Conference*, 2016.

[99] A. Sil and R. Florian, "One for all: Towards language independent named entity linking," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016. [Online]. Available: http://aclweb.org/anthology/P16-1213 pp. 2255–2264.

[100] G. Emerson, L. Tan, S. Fertmann, A. Palmer, and M. Regneri, "Seedling: Building and using a seed corpus for the human language project," in *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2014.

[101] S. Upadhyay, N. Gupta, and D. Roth, "Joint multilingual supervision for cross-lingual entity linking," in *Proc. EMNLP2018*, 2018.

[102] A. Sil, G. Kundu, R. Florian, and W. Hamza, "Neural cross-lingual entity linking," in *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[103] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of wikipedia entities in web text," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557073 pp. 457–466.

[104] M. Pennacchiotti and P. Pantel, "Entity extraction via ensemble semantics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2009. [Online]. Available: http://aclweb.org/anthology/D09-1025 pp. 238–247.

[105] N. Fernandez, J. A. Fisteus, L. Sanchez, and E. Martin, "Webtlab: A cooccurence-based approach to kbp 2010 entity-linking task," in *Proc. Text Analysis Conference (TAC 2010)*, 2010.

[106] W. Radford, B. Hachey, J. Nothman, M. Honnibal, and J. R. Curran, "Cmcrc at tac10: Document-level entity linking with graph-based re-ranking," in *Proc. Text Analysis Conference (TAC 2010)*, 2010.

[107] S. Cucerzan, "Tac entity linking by performing full-document entity extraction and disambiguation," in *Proc. Text Analysis Conference (TAC 2011)*, 2011.

[108] Y. Guo, W. Che, T. Liu, and S. Li, "A graph-based method for entity linking," in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 2011. [Online]. Available: http://aclweb.org/anthology/I11-1113 pp. 1010–1018.

[109] X. Han and L. Sun, "A generative entity-mention model for linking entities with knowledge base," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011. [Online]. Available: http://aclweb.org/anthology/P11-1095 pp. 945–954.

[110] J. Dalton and L. Dietz, "A neighborhood relevance model for entity linking," in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, ser. OAIR '13. Paris, France, France: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2491748.2491781 pp. 149–156.

[111] A. Chisholm and B. Hachey, "Entity disambiguation with web links," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 145–156, 2015. [Online]. Available: http://aclweb.org/anthology/Q15-1011

[112] H. Huang, Y. Cao, X. Huang, H. Ji, and C.-Y. Lin, "Collective tweet wikification based on semi-supervised graph regularization," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/P14-1036 pp. 380–390.

[113] H. Nguyen, H. Minha, T. Cao, and T. Nguyenb, "Jvn-tdt entity linking systems at tac-kbp2012," in *Proc. Text Analysis Conference (TAC 2012)*, 2012.

[114] J. Xu, Q. Lu, J. Liu, and R. Xu, "Nlpcomp in tac 2012 entity linking and slot-filling," in *Proc. Text Analysis Conference (TAC 2012)*, 2012.

[115] H. Huang, Y. Cao, X. Huang, H. Ji, and C.-Y. Lin, "Collective tweet wikification based on semi-supervised graph regularization," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/P14-1036 pp. 380–390.

[116] X. Ling, S. Singh, and D. S. Weld, "Context representation for named entity linking," in *Proc. Pacific Northwest Regional NLP Workshop (NW-NLP 2014)*, 2014.

[117] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/E14-1049 pp. 462–471.

[118] C. Xing, D. Wang, C. Liu, and Y. Lin, "Normalized word embedding and orthogonal transform for bilingual word translation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/N15-1104 pp. 1006–1011.

[119] A. Lazaridou, G. Dinu, and M. Baroni, "Hubness and pollution: Delving into cross-space mapping for zero-shot learning," 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)*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/P15-1027 pp. 270–280.

[120] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith, "Massively multilingual word embeddings," *CoRR*, vol. abs/1602.01925, 2016. [Online]. Available: http://arxiv.org/abs/1602.01925

[121] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017. [Online]. Available: http://aclweb.org/anthology/P17-1042 pp. 451–462.

[122] M. Zhang, Y. Liu, H. Luan, and M. Sun, "Adversarial training for unsupervised bilingual lexicon induction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017. [Online]. Available: http://aclweb.org/anthology/P17-1179 pp. 1959–1970.

[123] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. [Online]. Available: http://aclweb.org/anthology/P18-1073 pp. 789–798.

[124] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/D13-1141 pp. 1393–1398.

[125] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing crosslingual distributed representations of words," in *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, 2012. [Online]. Available: http://aclweb.org/anthology/C12-1089 pp. 1459–1474.

[126] T. Luong, H. Pham, and C. D. Manning, "Bilingual word representations with monolingual quality in mind," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/W15-1521 pp. 151–159.

[127] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith, "Massively multilingual word embeddings," *CoRR*, vol. abs/1602.01925, 2016. [Online]. Available: http://arxiv.org/abs/1602.01925

[128] I. Vulić, N. Mrkšić, and A. Korhonen, "Cross-lingual induction and transfer of verb classes based on word vector space specialisation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. [Online]. Available: http://aclweb.org/anthology/D17-1270 pp. 2546–2558.

[129] Y. Cao, L. Hou, J. Li, Z. Liu, C. Li, X. Chen, and T. Dong, "Joint representation learning of cross-lingual words and entities via attentive distant supervision," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. [Online]. Available: http://aclweb.org/anthology/D18-1021 pp. 227–237.

[130] J. Nothman, R. J. Curran, and T. Murphy, "Transforming wikipedia into named entity training data," in *Proceedings of the Australasian Language Technology Association Workshop 2008*, 2008. [Online]. Available: http://aclweb.org/anthology/U08-1016 pp. 124–132.

[131] W. Dakka and S. Cucerzan, "Augmenting wikipedia with named entity tags," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008. [Online]. Available: http://aclweb.org/anthology/I08-1071

[132] P. Mika, M. Ciaramita, H. Zaragoza, and J. Atserias, "Learning to tag and tagging to learn: A case study on wikipedia," *IEEE Intelligent Systems*, 2008.

[133] N. Ringland, J. Nothman, T. Murphy, and R. J. Curran, "Classifying articles in english and german wikipedia," in *Proceedings of the Australasian Language Technology Association Workshop 2009*, 2009. [Online]. Available: http://aclweb.org/anthology/U09-1004 pp. 20–28.

[134] F. Alotaibi and M. Lee, "Mapping arabic wikipedia into the named entities taxonomy," in *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, 2012. [Online]. Available: http://aclweb.org/anthology/C12-2005 pp. 43–52.

[135] M. Althobaiti, U. Kruschwitz, and M. Poesio, "Automatic creation of arabic named entity annotated corpus using wikipedia," in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014. [Online]. Available: http://aclweb.org/anthology/E14-3012 pp. 106–115.

[136] H. Ji and R. Grishman, "Analysis and repair of name tagger errors," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics, 2006. [Online]. Available: http://aclweb.org/anthology/P06-2055 pp. 420–427.

[137] E. A. Richman and P. Schone, "Mining wiki resources for multilingual named entity recognition," in *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, 2008. [Online]. Available: http://aclweb.org/anthology/P08-1001 pp. 1–9.

[138] M. Althobaiti, U. Kruschwitz, and M. Poesio, "A semi-supervised learning approach to arabic named entity recognition," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. INCOMA Ltd. Shoumen, BULGARIA, 2013. [Online]. Available: http://aclweb.org/anthology/R13-1005 pp. 32–40.

[139] M. Fleischman and E. Hovy, "Fine grained classification of named entities," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. [Online]. Available: http://aclweb.org/anthology/C02-1130

[140] C. Giuliano, "Fine-grained classification of named entities exploiting latent semantic kernels," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Association for Computational Linguistics, 2009. [Online]. Available: http://aclweb.org/anthology/W09-1125 pp. 201–209.

[141] A. Ekbal, E. Sourjikova, A. Frank, and S. P. Ponzetto, "Assessing the challenge of fine-grained named entity recognition and classification," in *Proceedings of the 2010 Named Entities Workshop*. Association for Computational Linguistics, 2010. [Online]. Available: http://aclweb.org/anthology/W10-2415 pp. 93–101.

[142] X. Ling and D. S. Weld, "Fine-grained entity recognition," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, pp. 94–100.

[143] A. M. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum, "Hyena: Hierarchical type classification for entity names," in *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, 2012. [Online]. Available: http://aclweb.org/anthology/C12-2133 pp. 1361–1370.

[144] N. Nakashole, T. Tylenda, and G. Weikum, "Fine-grained semantic typing of emerging entities," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/P13-1146 pp. 1488–1497.

[145] D. Gillick, N. Lazic, K. Ganchev, J. Kirchner, and D. Huynh, "Context-dependent fine-grained entity type tagging," *CoRR*, vol. abs/1412.1820, 2014. [Online]. Available: http://arxiv.org/abs/1412.1820

[146] D. Yogatama, D. Gillick, and N. Lazic, "Embedding methods for fine grained entity type classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/P15-2048 pp. 291–296.

[147] L. Del Corro, A. Abujabal, R. Gemulla, and G. Weikum, "Finet: Context-aware fine-grained named entity typing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/D15-1103 pp. 868–878.

[148] Y. Yaghoobzadeh and H. Schütze, "Corpus-level fine-grained entity typing using contextual information," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015. [Online]. Available: http://aclweb.org/anthology/D15-1083 pp. 715–725.

[149] Y. Ma, E. Cambria, and S. GAO, "Label embedding for zero-shot fine-grained named entity typing," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016. [Online]. Available: http://aclweb.org/anthology/C16-1017 pp. 171–180.

[150] J. Kazama and K. Torisawa, "Exploiting wikipedia as external knowledge for named entity recognition," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007. [Online]. Available: http://aclweb.org/anthology/D07-1073

[151] S. Paolo Ponzetto and M. Strube, "Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 2006. [Online]. Available: http://aclweb.org/anthology/N06-1025

[152] S. Y. Chan and D. Roth, "Exploiting background knowledge for relation extraction," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, 2010. [Online]. Available: http://aclweb.org/anthology/C10-1018 pp. 152–160.

[153] A. Hogue, J. Nothman, and J. R. Curran, "Unsupervised biographical event extraction using wikipedia traffic," in *Proceedings of the Australasian Language Technology Association Workshop 2014*, 2014. [Online]. Available: http://aclweb.org/anthology/U14-1006 pp. 41–49.

[154] J. An, S. Lee, and G. Geunbae Lee, "Automatic acquisition of named entity tagged corpus from world wide web," in *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, 2003. [Online]. Available: http://aclweb.org/anthology/P03-2031

[155] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 2009. [Online]. Available: http://aclweb.org/anthology/P09-1113 pp. 1003–1011.

[156] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, and J. Han, "Clustype: Effective entity recognition and typing by relation phrase-based clustering," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2783258.2783362 pp. 995–1004.

[157] Q. Li, H. Li, H. Ji, W. Wang, J. Zheng, and F. Huang, "Joint bilingual name tagging for parallel corpora," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2396761.2398506 pp. 1727–1731.

[158] S. Kim, K. Toutanova, and H. Yu, "Multilingual named entity recognition using parallel data and metadata from wikipedia," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2012. [Online]. Available: http://aclweb.org/anthology/P12-1073 pp. 694–702.

[159] W. Che, M. Wang, D. C. Manning, and T. Liu, "Named entity recognition with bilingual constraints," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/N13-1006 pp. 52–62.

[160] M. Wang, W. Che, and D. C. Manning, "Joint word alignment and bilingual named entity recognition using dual decomposition," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/P13-1106 pp. 1073–1082.

[161] M. Wang and D. C. Manning, "Cross-lingual projected expectation regularization for weakly supervised learning," *Transactions of the Association of Computational Linguistics*, vol. 2, pp. 55–66, 2014. [Online]. Available: http://aclweb.org/anthology/Q14-1005

[162] D. Zhang, B. Zhang, X. Pan, X. Feng, H. Ji, and W. XU, "Bitext name tagging for cross-lingual entity annotation projection," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016. [Online]. Available: http://aclweb.org/anthology/C16-1045 pp. 461–470.

[163] P. Littell, K. Goyal, R. D. Mortensen, A. Little, C. Dyer, and L. Levin, "Named entity recognition for linguistic rapid response in low-resource languages: Sorani kurdish and tajik," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016. [Online]. Available: http://aclweb.org/anthology/C16-1095 pp. 998–1006.

[164] Y. Lin, S. Yang, V. Stoyanov, and H. Ji, "A multi-lingual multi-task architecture for low-resource sequence labeling," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. [Online]. Available: http://aclweb.org/anthology/P18-1074 pp. 799–809.

[165] L. Huang, K. Cho, B. Zhang, H. Ji, and K. Knight, "Multi-lingual common semantic space construction via cluster-consistent word embedding," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. [Online]. Available: http://aclweb.org/anthology/D18-1023 pp. 250–260.

[166] U. Hermjakob, J. May, M. Pust, and K. Knight, "Translating a language you don't know in the chinese room," in *Proceedings of ACL2018 Demo Track*, 2018.

[167] Y. Lin, C. Costello, B. Zhang, D. Lu, H. Ji, J. Mayfield, and P. McNamee, "Platforms for non-speakers annotating names in any language," in *Proceedings of ACL2018 Demo Track*, 2018.

[168] S. Mayhew and D. Roth, "Talen: Tool for annotation of low-resource entities," in *Proceedings of ACL2018 Demo Track*, 2018.

[169] M. Kobayashi, A. Ishii, C. Hoshino, H. Miyashita, and T. Matsuzaki, "Automated historical fact-checking by passage retrieval, word statistics, and virtual question-answering," in *Proceedings of the IJCNLP*, Taipei, Taiwan, 2017. [Online]. Available: http://www.aclweb.org/anthology/I17-1097

[170] J. Welbl, N. F. Liu, and M. Gardner, "Crowdsourcing multiple choice science questions," in *Proceedings of the W-NUT*, Copenhagen, Denmark, 2017. [Online]. Available: http://www.aclweb.org/anthology/W17-4413

[171] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," *arXiv preprint*, vol. cs.CL/1611.09268v2, 2016. [Online]. Available: https://arxiv.org/abs/1611.09268v2

[172] B. Dhingra, K. Mazaitis, and W. W. Cohen, "Quasar: Datasets for question answering by search and reading," *arXiv preprint*, vol. cs.CL/1707.03904v2, 2017. [Online]. Available: https://arxiv.org/abs/1707.03904v2

[173] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint*, vol. cs.CL/1705.03551v2, 2017. [Online]. Available: http://arxiv.org/abs/1705.03551v2

[174] M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho, "SearchQA: A new Q&A dataset augmented with context from a search engine," *arXiv preprint*, vol. cs.CL/1704.05179v3, 2017. [Online]. Available: https://arxiv.org/abs/1704.05179v3

[175] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: A benchmark for question answering research," *TACL*, 2019. [Online]. Available: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/b8c26e4347adc3453c15d96a09e6f7f102293f71.pdf

[176] M. Hu, Y. Peng, Z. Huang, N. Yang, M. Zhou et al., "Read+Verify: Machine reading comprehension with unanswerable questions," in *Proceedings of the AAAI*, Honolulu, HI, 2019. [Online]. Available: https://arxiv.org/abs/1808.05759v5

[177] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, "End-to-end open-domain question answering with bertserini," *arXiv preprint*, vol. cs.CL/1902.01718v1, 2019. [Online]. Available: https://arxiv.org/abs/1902.01718v1

[178] H. Kwon, H. Trivedi, P. Jansen, M. Surdeanu, and N. Balasubramanian, "Controlling information aggregation for complex question answering," in *Proceedings of the ECIR*, Grenoble, France, 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-76941-7_72

[179] D. Khashabi, T. Khot, A. Sabharwal, and D. Roth, "Question answering as global reasoning over semantic abstractions," in *Proceedings of the AAAI*, New Orleans, LA, 2018. [Online]. Available: http://www.cis.upenn.edu/~danielkh/files/2018_semanticilp/2018_aaai_semanticilp.pdf

[180] Y. Li and P. Clark, "Answering elementary science questions by constructing coherent scenes using background knowledge," in *Proceedings of the EMNLP*, Lisbon, Portugal, 2015. [Online]. Available: http://www.aclweb.org/anthology/D15-1236

[181] M. Sachan, A. Dubey, and E. P. Xing, "Science question answering using instructional materials," in *Proceedings of the ACL*, Berlin, Germany, 2016. [Online]. Available: http://www.aclweb.org/anthology/P16-2076

[182] X. Wang, P. Kapanipathi, R. Musa, M. Yu, K. Talamadupula, I. Abdelaziz, M. Chang, A. Fokoue, B. Makni, N. Mattei et al., "Improving natural language inference using external knowledge in the science questions domain," *arXiv preprint*, vol. cs.CL/1809.05724v2, 2018. [Online]. Available: https://arxiv.org/abs/1809.05724v2

[183] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI*, San Francisco, CA, 2017. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPDFInterstitial/14972/14051

[184] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the EMNLP*, Austin, TX, 2016. [Online]. Available: http://www.aclweb.org/anthology/D16-1264

[185] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the ACL*, Vancouver, Canada, 2017. [Online]. Available: http://www.aclweb.org/anthology/P17-1171

[186] S. Wang, M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, and J. Jiang, "R^3: Reinforced reader-ranker for open-domain question answering," in *Proceedings of the AAAI*, New Orleans, LA, 2018. [Online]. Available: https://arxiv.org/abs/1709.00023v2

[187] B. Kratzwald and S. Feuerriegel, "Adaptive document retrieval for deep question answering," in *Proceedings of the EMNLP*, Brussels, Belgium, 2018. [Online]. Available: http://www.aclweb.org/anthology/D18-1055

[188] J. Lee, S. Yun, H. Kim, M. Ko, and J. Kang, "Ranking paragraphs for improving answer recall in open-domain question answering," in *Proceedings of the EMNLP*, Brussels, Belgium, 2018. [Online]. Available: http://www.aclweb.org/anthology/D18-1053

[189] Y. Lin, H. Ji, Z. Liu, and M. Sun, "Denoising distantly supervised open-domain question answering," in *Proceedings of the ACL*, Melbourne, Australia, 2018. [Online]. Available: http://www.aclweb.org/anthology/P18-1161

[190] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "QANet: Combining local convolution with global self-attention for reading comprehension," in *Proceedings of the ICLR*, Vancouver, Canada, 2018. [Online]. Available: https://arxiv.org/abs/1804.09541v1