Cross-lingual Name Tagging and Linking for 282 Languages

Xiaoman Pan[1], Boliang Zhang[1], Jonathan May[2], Joel Nothman[3], Kevin Knight[2], Heng Ji[1]

[1] Computer Science Department, Rensselaer Polytechnic Institute
{panx2,zhangb8,jih}@rpi.edu

[2] Information Sciences Institute, University of Southern California
{jonmay,knight}@isi.edu

[3] Sydney Informatics Hub, University of Sydney
joel.nothman@gmail.com

Abstract

The ambitious goal of this work is to develop a cross-lingual name tagging and linking framework for 282 languages that exist in Wikipedia. Given a document in any of these languages, our framework is able to identify name mentions, assign a coarse-grained or fine-grained type to each mention, and link it to an English Knowledge Base (KB) if it is linkable. We achieve this goal by performing a series of new KB mining methods: generating "silver-standard" annotations by transferring annotations from English to other languages through cross-lingual links and KB properties, refining annotations through self-training and topic selection, deriving language-specific morphology features from anchor links, and mining word translation pairs from cross-lingual links. Both name tagging and linking results for 282 languages are promising on Wikipedia data and on-Wikipedia data. All the data sets, resources and systems for 282 languages are made publicly available as a new benchmark [1].

1 Introduction

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 which is accessible to a user in his/her own language (e.g., a reporter from the World Health Organization who speaks English only).

Wikipedia is a massively multi-lingual resource that currently hosts 295 languages and contains naturally annotated markups [2] and rich informational 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[3], YAGO (Mahdisoltani et al., 2015) and Freebase (Bollacker et al., 2008) that contain rich properties. Figure 1 shows an example of Wikipedia markups and KB properties. We leverage these markups for developing a language universal framework to automatically extract name mentions from documents in

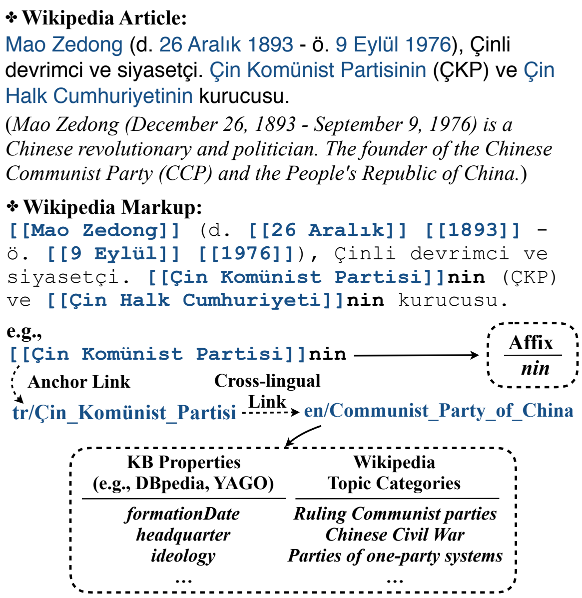

Figure 1: Examples of Wikipedia Markups and KB Properties

[1] http://nlp.cs.rpi.edu/wikiann

[2] https://en.wikipedia.org/wiki/Help:Wiki_markup

[3] http://wiki.dbpedia.org

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1946–1958
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1178

282 languages, and link them to an English KB (Wikipedia in this work). 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 (Banarescu et al., 2013) 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 (Section 2.2).

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 (Section 2.4).

Customize annotations through cross-lingual topic transfer. For the first time, we propose to customize name annotations for specific downstream 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 (Section 2.5).

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 (Roth et al., 2008; Mahmoudi et al., 2013; Ahlberg et al., 2015), or low-quality unsupervised learning (Grönroos et al., 2014; Ruokolainen et al., 2016). We exploit Wikipedia markups to automatically learn affixes as language-specific features (Section 2.3).

Mine word translations from cross-lingual links. Name translation is a crucial step to generate candidate entities in cross-lingual entity linking. Only a small percentage of names can be directly translated by matching against cross-lingual Wikipedia title pairs. Based on the observation that Wikipedia titles within any language tend to follow a consistent style and format, we propose an effective method to derive word translation pairs from these titles based on automatic alignment (Section 3.2).

2 Name Tagging

2.1 Overview

Our first step is to generate "silver-standard" name annotations from Wikipedia markups and train a universal name tagger. Figure 2 shows our overall procedure and the following subsections will elaborate each component.

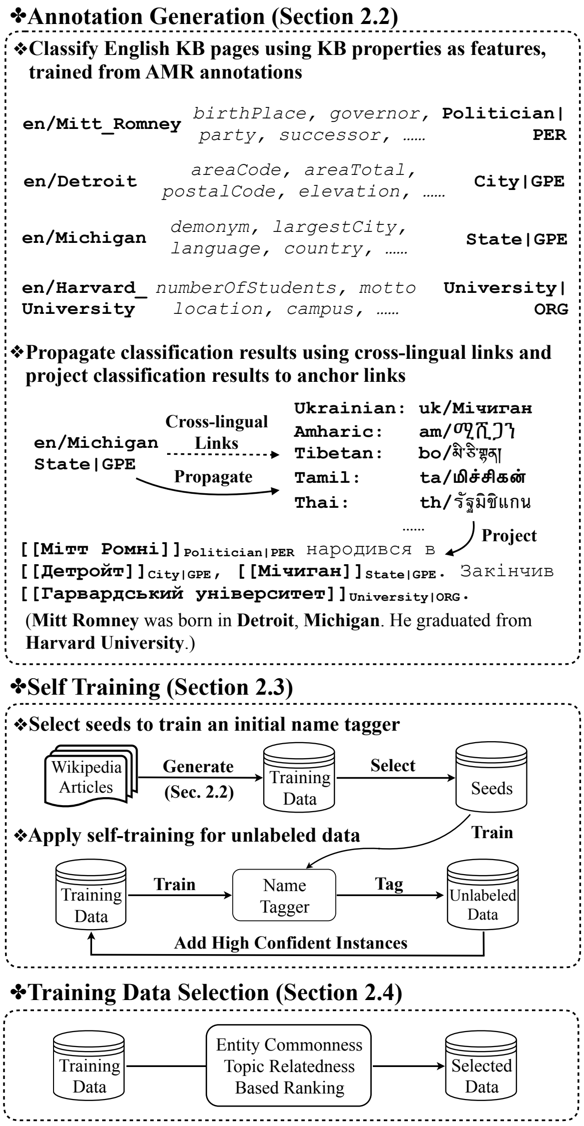

Figure 2: Name Tagging Annotation Generation and Training

2.2 Initial Annotation Generation

We 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) (Ward Church and Hanks, 1990) 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.

Nothman et al. (2013) 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 entity mentions in all languages in the entire Wikipedia through monolingual redirect links and cross-lingual links.

2.3 Learning Model and KB Derived Features

We use a typical neural network architecture that consists of Bi-directional Long Short-Term Memory and Conditional Random Fields (CRFs) network (Lample et al., 2016) as our underlying learning model for the name tagger for each language. In the following we will describe how we acquire linguistic features.

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.

Table 1 summarizes name tagging features.

Features	Descriptions
Form	Lowercase forms of (w_{-1}, w_0, w_{+1})
Case	Case of w_0
Syllable	The first and the last character of w_0
Stem	Stems of (w_{-1}, w_0, w_{+1})
Affix	Affixes of (w_{-1}, w_0, w_{+1})
Gazetteer	Cross-lingual gazetteers learned from training data
Embeddings	Character embeddings and word embeddings [4]learned from training data

Table 1: Name Tagging Features

2.4 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. We apply self-training to propagate and refine the labels.

We first train an initial name tagger using seeds selected from the labeled data. We adopt an idea from (Guo et al., 2014) which computes Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009) between a tag and a token:

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

$$NPMI(tag, token) = \frac{\ln \frac{p(tag,token)}{p(tag)p(token)}}{-\ln p(tag, token)} \quad (1)$$

Then we select the sentences in which all annotations satisfy $NPMI(tag, token) > \tau$ as seeds [5].

For all Wikipedia articles in a language, we cluster the unlabeled sentences into n clusters [6] 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.

2.5 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 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 [7], 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 (Li et al., 2011) found that by carefully select 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 (Manning et al., 2014) and entity linker (Pan et al., 2015) on the Leidos corpus released by the DARPA LORELEI program [8]. 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 final model.

[5] $\tau = 0$ in our experiment.
[6] $n = 20$ in our experiment.
[7] D = 30,000 in our experiment.

3 Cross-lingual Entity Linking

3.1 Overview

After we extract names from test documents in a source language, we translate them into English by automatically mined word translation pairs (Section 3.2), and then link translated English mentions to an external English KB (Section 3.3). The overall linking process is illustrated in Figure 3.

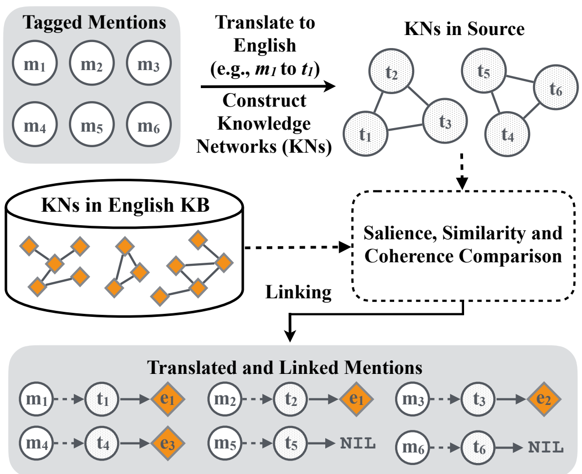

Figure 3: Cross-lingual Entity Linking Overview

3.2 Name Translation

The cross-lingual Wikipedia title pairs, generated through crowd-sourcing, generally follow a consistent style and format in each language. From Table 2 we can see that the order of modifier and head word keeps consistent in Turkish and English titles.

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

Extracted Cross-lingual Wikipedia Title Pairs	
"Pekin"	
Pekin	Beijing
Pekin metrosu	Beijing Subway
Pekin Ulusal Stadyumu	Beijing National Stadium
"Teknoloji"	
Nükleer **teknoloji**	Nuclear technology
Teknoloji transferi	Technology transfer
Teknoloji eğitimi	Technology education
"Enstitüsü"	
Torchwood **Enstitüsü**	Torchwood Institute
Hudson **Enstitüsü**	Hudson Institute
Smolny **Enstitüsü**	Smolny Institute
"Pekin Teknoloji" [NONE]	
"Teknoloji Enstitüsü"	
Kraliyet **Teknoloji Enstitüsü**	Royal Institute of Technology
Karlsruhe **Teknoloji Enstitüsü**	Karlsruhe Institute of Technology
Georgia **Teknoloji Enstitüsü**	Georgia Institute of Technology
"Pekin Teknoloji Enstitüsü" [NONE]	

Mined Word Translation Pairs		
Word	Translation	Alignment Confidence
pekin	**Beijing**	*Exact Match*
	beijing	0.5263
	peking	0.3158
teknoloji	**technology**	0.8833
	technological	0.0167
	singularity	0.0167
enstitüsü	**institute**	0.2765
	of	0.2028
	for	0.0221

Table 2: Word Translation Mining from Cross-lingual Wikipedia Title Pairs

For each name mention, we generate all possible combinations of continuous tokens. For example, no Wikipedia titles contain the Turkish name "*Pekin Teknoloji Enstitüsü (Beijing Institute of Technology)*". We generate the following 6 combinations: "*Pekin*", "*Teknoloji*", "*Enstitüsü*", "*Pekin Teknoloji*", "*Teknoloji Enstitüsü*" and "*Pekin Teknoloji Enstitüsü*", and then extract all cross-lingual Wikipedia title pairs containing each combination. Finally we run GIZA++ (Josef Och and Ney, 2003) to extract word for word translations from these title pairs, as shown in Table 2.

3.3 Entity Linking

Given a set of tagged name mentions $M = \{m_1, m_2, ..., m_n\}$, we first obtain their English translations $T = \{t_1, t_2, ..., t_n\}$ using the approach described above. Then we apply an unsupervised collective inference approach to link T to the KB, similar to our previous work (Pan et al., 2015). The only difference is that we construct knowledge networks (KNs) $g(t_i)$ for T based on their co-occurrence within a context window [9] instead of their AMR relations, because AMR parsing is not available for foreign languages. For each translated name mention t_i, an initial list of candidate entities $E(t_i) = \{e_1, e_2, ..., e_k\}$ is generated based on a surface form dictionary mined from KB properties (e.g., *redirects*, *names*, *aliases*). If no surface form can be matched then we determine the mention as unlinkable. Then we construct KNs $g(e_j)$ for each entity candidate e_j in t_i's entity candidate list $E(t_i)$. We compute the similarity between $g(t_i)$ and $g(e_j)$ based on three measures: salience, similarity and coherence, and select the candidate entity with the highest score.

4 Experiments

4.1 Performance on Wikipedia Data

We first conduct an evaluation using Wikipedia data as "silver-standard". For each language, we use 70% of the selected sentences for training and 30% for testing. For entity linking, we don't have ground truth for unlinkable mentions, so we only compute linking accuracy for linkable name mentions. Table 3 presents the overall performance for three coarse-grained entity types: PER, ORG and GPE/LOC, sorted by the number of name mentions. Figure 4 and Figure 5 summarize the performance, with some example languages marked for various ranges of data size.

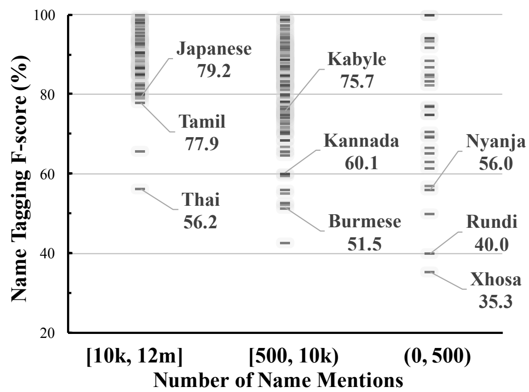

Figure 4: Summary of Name Tagging F-score (%) on Wikipedia Data

Not surprisingly, name tagging performs better for languages with more training mentions. The

[9] In our experiments, we use the previous four and next four name mentions as a context window.

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 name tagging 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.

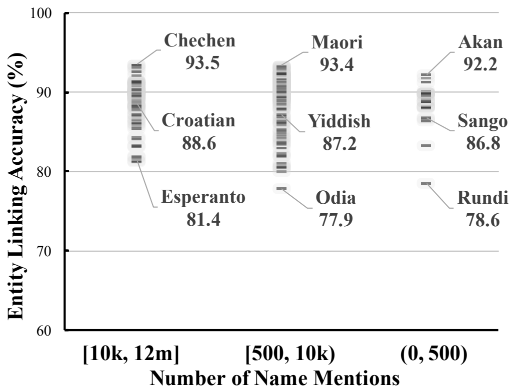

Figure 5: Summary of Entity Linking Accuracy (%) on Wikipedia Data

The linking accuracy is higher than 80% 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 are not directly comparable to traditional name tagging results measured against gold-standard data annotated by human.

[10] The mapping to language names can be found at http://nlp.cs.rpi.edu/wikiann/mapping

4.2 Performance on Non-Wikipedia Data

In order to have more direct comparison with state-of-the-art name taggers 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 shows the data statistics. The documents are from news sources and discussion fora.

For fair comparison, we use the same learning method and feature set as described in Section 2.3 to train the models using gold-standard data. Therefore the results of our models trained from gold-standard data are slightly different from some previous work such as (Tsai et al., 2016), mainly due to different learning algorithms and different features sets. For example, the gazetteers we used are different from those in (Tsai et al., 2016), and we did not use brown clusters as additional features.

The name tagging results on LORELEI data set are presented in Table 5. We can see that our approach advances state-of-the-art language-independent methods (Zhang et al., 2016a; Tsai et al., 2016) 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.

Mayfield et al. (2011) constructed a cross-lingual entity linking collection 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 (McNamee et al., 2011). The entity linking results on non-NIL mentions are presented in Table 6. 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.

4.3 Analysis

Impact of KB-derived Morphological Features

We measured the impact of our affix lists derived from Wikipedia markups on two morphologically-rich languages: Turkish and Uzbek. The morphol-

[11] McNamee et al. (2011) did not develop a model for Chinese even though Chinese data set was included in the collection.

L	*M*	*F*	*A*	*L*	*M*	*F*	*A*	*L*	*M*	*F*	*A*	*L*	*M*	*F*	*A*
en	12M	91.8	84.3	mr	18K	82.4	89.8	szl	3.0K	82.7	92.2	tet	1.2K	73.5	92.2
ja	1.9M	79.2	86.7	bar	17K	97.1	93.1	tk	2.9K	86.3	90.1	sc	1.2K	78.1	91.6
sv	1.8M	93.6	89.7	cv	15K	95.7	93.2	z-c	2.9K	88.2	87.0	wuu	1.2K	79.7	90.8
de	1.7M	89.0	89.8	ba	15K	93.8	92.6	mn	2.9K	76.4	84.4	ksh	1.2K	56.0	83.6
fr	1.4M	93.3	91.2	mg	14K	98.7	90.1	kv	2.9K	89.7	93.2	pfl	1.1K	42.9	80.4
ru	1.4M	90.1	90.0	hi	14K	86.9	88.0	f-v	2.9K	65.4	88.8	haw	1.1K	88.0	84.6
it	1.2M	96.6	90.2	an	14K	93.0	91.1	gan	2.9K	84.9	90.9	am	1.1K	84.7	83.0
sh	1.1M	97.8	90.9	als	14K	85.0	90.9	fur	2.8K	84.5	89.2	bcl	1.1K	82.3	91.7
es	992K	93.9	90.2	sco	14K	86.8	89.6	kw	2.8K	94.0	93.3	nah	1.1K	89.9	89.6
pl	931K	90.0	91.3	bug	13K	99.9	90.0	ilo	2.8K	90.3	91.1	udm	1.1K	88.9	85.0
nl	801K	93.2	91.5	lb	13K	81.5	88.4	mwl	2.7K	76.1	89.4	su	1.1K	72.7	89.2
zh	718K	82.0	90.0	fy	13K	86.6	91.2	mai	2.7K	99.7	90.0	dsb	1.1K	84.7	82.1
pt	576K	90.7	90.3	new	12K	98.2	91.5	nv	2.7K	90.9	91.6	tpi	1.1K	83.3	90.1
uk	472K	91.5	89.4	ga	12K	85.3	91.3	sd	2.7K	65.8	90.9	lo	1.0K	52.8	88.6
cs	380K	94.6	90.5	ht	12K	98.9	93.4	os	2.7K	87.4	89.4	bpy	1.0K	98.3	89.3
sr	365K	95.3	91.2	war	12K	94.9	89.8	mzn	2.6K	86.4	86.9	ki	1.0K	97.5	90.0
hu	357K	95.9	90.4	te	11K	80.5	86.1	azb	2.6K	88.4	90.6	ty	1.0K	86.7	89.8
fi	341K	93.4	90.6	is	11K	80.2	83.2	bxr	2.6K	75.0	90.3	hif	1.0K	81.1	93.1
no	338K	94.1	90.6	pms	10K	98.0	89.5	vec	2.6K	87.9	91.3	ady	979	92.7	91.2
fa	294K	96.4	86.4	zea	10K	86.8	90.3	bo	2.6K	70.4	88.9	ig	968	74.4	91.8
ko	273K	90.6	89.8	sw	9.3K	93.4	90.8	yi	2.6K	76.9	87.2	tyv	903	91.1	91.0
ca	265K	90.3	90.3	ia	8.9K	75.4	90.5	frp	2.5K	86.2	92.3	tn	902	76.9	90.1
tr	223K	96.9	87.3	qu	8.7K	92.5	88.2	myv	2.5K	88.6	92.2	cu	898	75.5	91.3
ro	197K	90.6	89.2	ast	8.3K	89.2	92.0	se	2.5K	90.3	83.5	sm	888	80.0	85.3
bg	186K	65.8	88.4	rm	8.0K	82.0	91.3	cdo	2.5K	91.0	91.9	to	866	92.3	90.7
ar	185K	88.3	89.7	ay	7.9K	88.5	91.0	nso	2.5K	98.9	90.0	tum	831	93.8	92.9
id	150K	87.8	90.0	ps	7.7K	66.9	89.9	gom	2.4K	88.8	90.0	r-r	750	93.0	85.9
he	145K	79.0	91.0	mi	7.5K	95.9	93.4	ky	2.4K	71.8	88.4	om	709	74.2	81.1
eu	137K	82.5	89.2	gag	7.3K	89.3	84.0	n-n	2.3K	92.6	91.6	glk	688	59.5	80.7
da	133K	87.1	85.8	nds	7.0K	84.5	89.8	ne	2.3K	81.5	91.1	lbe	651	88.9	90.8
vi	125K	89.6	82.0	gd	6.7K	92.8	91.3	sa	2.2K	73.9	91.3	bjn	640	64.7	89.5
th	96K	56.2	87.7	mrj	6.7K	97.0	91.6	mt	2.2K	82.3	90.3	srn	619	76.5	89.3
sk	93K	87.3	90.3	so	6.5K	85.8	91.7	my	2.2K	51.5	91.2	mdf	617	82.2	92.4
uz	92K	98.3	90.3	co	6.0K	85.4	89.9	bh	2.2K	92.6	92.5	tw	572	94.6	90.4
eo	85K	88.7	81.4	pnb	6.0K	90.8	86.2	vls	2.2K	78.2	89.1	pih	555	87.2	89.0
la	81K	90.8	89.4	pcd	5.8K	86.1	90.8	ug	2.1K	79.7	92.4	rmy	551	68.5	86.4
z-m	79K	99.3	89.2	wa	5.8K	81.6	82.0	si	2.1K	87.7	90.5	lg	530	98.8	89.3
lt	79K	86.3	87.2	frr	5.7K	70.1	86.3	kaa	2.1K	55.2	89.5	chr	530	70.6	86.2
el	78K	84.6	88.3	scn	5.6K	93.2	89.2	b-s	2.1K	84.5	88.0	ha	517	75.0	87.9
ce	77K	99.4	93.5	fo	5.4K	83.6	92.2	krc	2.1K	84.9	88.9	ab	506	60.0	92.4
ur	77K	96.4	89.3	ckb	5.3K	88.1	89.3	ie	2.1K	88.8	92.8	got	506	91.7	90.1
hr	76K	82.8	88.5	li	5.2K	89.4	91.3	dv	2.0K	76.2	90.5	bi	490	88.5	88.3
ms	75K	86.8	84.1	nap	4.9K	86.9	89.9	xmf	2.0K	73.4	92.2	st	455	84.4	89.8
et	69K	86.8	89.9	crh	4.9K	90.1	89.9	rue	1.9K	82.7	92.2	chy	450	85.1	89.9
kk	68K	88.3	81.8	gu	4.6K	76.0	90.8	pa	1.8K	74.8	84.3	iu	450	66.7	88.9
ceb	68K	96.3	86.6	km	4.6K	52.2	89.9	eml	1.8K	83.5	88.5	zu	449	82.3	89.9
sl	67K	89.5	90.1	tg	4.5K	88.3	90.6	arc	1.8K	68.5	89.2	pnt	445	61.5	89.6
nn	65K	88.1	89.9	hsb	4.5K	91.5	92.0	pdc	1.8K	78.1	91.1	ik	436	94.1	88.2
sim	59K	85.7	90.7	c-z	4.5K	75.0	86.6	kbd	1.7K	74.9	80.6	lrc	416	65.2	86.9
lv	57K	92.1	89.8	jv	4.4K	82.6	87.8	pap	1.7K	88.8	58.4	bm	386	77.3	89.1
tt	53K	87.7	91.4	lez	4.4K	84.2	82.3	jbo	1.7K	92.4	91.6	za	382	57.1	88.2
gl	52K	87.4	88.2	hak	4.3K	85.5	88.1	diq	1.7K	79.3	80.9	mo	373	69.6	88.2
ka	49K	79.8	89.5	ang	4.2K	84.0	92.0	pag	1.7K	91.2	89.5	ss	362	69.2	91.8
vo	47K	98.5	90.8	r-t	4.2K	88.1	89.0	kg	1.6K	82.1	90.1	ee	297	63.2	90.0
lmo	39K	98.3	89.0	kn	4.1K	60.1	91.7	m-b	1.6K	78.3	80.0	dz	262	50.0	90.0
be	38K	84.1	88.3	csb	4.1K	87.0	92.3	rw	1.6K	95.4	91.5	ak	258	86.8	92.2
mk	35K	93.4	83.3	lij	4.1K	72.3	91.9	or	1.6K	86.4	77.9	sg	245	99.9	86.8
cy	32K	90.7	89.3	nov	4.0K	77.0	92.1	ln	1.6K	82.8	91.4	ts	236	93.3	88.9
bs	31K	84.8	89.8	ace	4.0K	81.6	90.3	kl	1.5K	75.0	90.9	rn	185	40.0	78.6
ta	31K	77.9	88.2	gn	4.0K	71.2	89.3	sn	1.5K	95.0	93.3	ve	183	99.9	88.0
hy	28K	90.4	81.3	koi	4.0K	89.6	92.9	av	1.4K	82.0	83.7	ny	169	56.0	90.2
bn	27K	93.8	87.2	mhr	3.9K	86.7	92.4	as	1.4K	89.6	89.3	ff	168	76.9	88.9
az	26K	85.1	86.0	io	3.8K	87.2	92.3	stq	1.4K	70.0	90.6	ch	159	70.6	90.0
sq	26K	94.1	92.1	min	3.8K	85.8	89.9	gv	1.3K	84.8	89.1	xh	141	35.3	89.5
ml	24K	82.4	88.8	arz	3.8K	77.8	89.3	wo	1.3K	87.7	90.0	fj	126	75.0	91.3
br	22K	87.0	85.5	ext	3.7K	77.8	91.6	xal	1.3K	98.7	90.9	ks	124	75.0	83.3
z-y	22K	87.3	88.4	yo	3.7K	94.0	90.8	nrm	1.3K	96.4	92.7	ti	52	94.2	90.0
af	21K	85.7	91.1	sah	3.6K	91.2	93.0	na	1.2K	87.6	88.7	cr	49	91.8	89.8
b-x	20K	85.1	87.7	vep	3.5K	85.8	89.8	ltg	1.2K	74.3	92.1	pi	41	83.3	86.4
tl	19K	92.7	90.3	ku	3.3K	83.2	85.1	pam	1.2K	87.2	91.0				
oc	18K	92.5	90.0	kab	3.3K	75.7	84.3	lad	1.2K	92.3	92.4				

Table 3: Performance on Wikipedia Data (*L*: language ID [10]; *M*: the number of name mentions; *F*: name tagging F-score (%); *A*: entity linking accuracy (%))

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: # of Names in Non-Wikipedia Data

Language	Training from *Gold*	Training from *Silver*	(Zhang et al., 2016a)	(Tsai et al., 2016)
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 5: Name Tagging F-score (%) on Non-Wikipedia Data

Language	# of Non-NIL Mentions	(Mayfield et al., 2011)	Our Approach
Arabic	661	70.6	**80.2**
Bulgarian	2,068	82.1	**84.1**
Chinese	956	- [11]	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 6: Entity Linking Accuracy (%) on Non-Wikipedia Data

ogy features contributed 11.1% and 7.1% absolute name tagging F-score gains to Turkish and Uzbek LORELEI data sets respectively.

Impact of Self-Training

Using Turkish as a case study, the learning curves of self-training on Wikipedia and non-Wikipedia test sets are shown in Figure 6. We can see that self-training provides significant improvement for both Wikipedia (6% absolute gain) and non-Wikipedia test data (12% absolute gain). As expected the learning curve on Wikipedia data is more smooth and converges more slowly than that of non-Wikipedia data. This indicates that when the training data is incomplete and noisy, the model can benefit from self-training through iterative label correction and propagation.

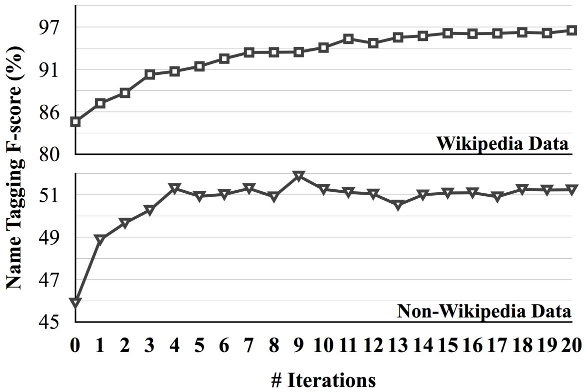

Figure 6: Learning Curve of Self-training

Impact of Topical Relatedness

We also found that the topical relatedness measure proposed in Section 2.5 not only significantly reduces the size of training data and thus speeds up the training process for many languages, but also consistently improves the quality. For example, the Turkish name tagger trained from the entire data set without topic selection yields 49.7% F-score on LORELEI data set, and the performance is improved to 51.5% after topic selection.

5 Related Work

Wikipedia markup based silver standard generation: Our work was mainly inspired from previous work that leveraged Wikipedia markups to train name taggers (Nothman et al., 2008; Dakka and Cucerzan, 2008; Mika et al., 2008; Ringland et al., 2009; Alotaibi and Lee, 2012; Nothman et al., 2013; Althobaiti et al., 2014). Most of these previous methods manually classified 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.

Many fine-grained entity typing approaches (Fleischman and Hovy, 2002; Giuliano, 2009; Ekbal et al., 2010; Ling and Weld, 2012; Yosef et al., 2012; Nakashole et al., 2013; Gillick et al., 2014; Yogatama et al., 2015; Del Corro et al., 2015) also created annotations based on Wikipedia anchor links. Our framework performs both name identification and typing and takes advantage of richer structures in the KBs. Previous work on Arabic name tagging (Althobaiti et al., 2014) extracted entity titles as a gazetteer for stemming, and thus it cannot handle unknown names. We developed a new method to derive generalizable affixes for morphologically rich language based on Wikipedia markups.

Wikipedia as background features for IE: Wikipedia pages have been used as additional features to improve various Information Extraction (IE) tasks, including name tagging (Kazama and Torisawa, 2007), coreference resolution (Paolo Ponzetto and Strube, 2006), relation extraction (Chan and Roth, 2010) and event extraction (Hogue et al., 2014). Other automatic name annotation generation methods have been proposed, including KB driven distant supervision (An et al., 2003; Mintz et al., 2009; Ren et al., 2015) and cross-lingual projection (Li et al., 2012; Kim et al., 2012; Che et al., 2013; Wang et al., 2013; Wang and Manning, 2014; Zhang et al., 2016b).

Multi-lingual name tagging: Some recent research (Zhang et al., 2016a; Littell et al., 2016; Tsai et al., 2016) under the DARPA LORELEI program focused on developing name tagging techniques for low-resource languages. These approaches require English annotations for projection (Tsai et al., 2016), some input from a native speaker, either through manual annotations (Littell et al., 2016), or a linguistic survey (Zhang et al., 2016a). Without using any manual annotations, our name taggers outperform previous methods on the same data sets for many languages.

Multi-lingual entity linking: NIST TAC-KBP Tri-lingual entity linking (Ji et al., 2016) focused on three languages: English, Chinese and Spanish. (McNamee et al., 2011) extended it to 21 languages. But their methods required labeled data and name transliteration. We share the same goal as (Sil and Florian, 2016) 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 (Pan et al., 2015) to directly link translated English name mentions to English KB.

Efforts to save annotation cost for name tagging: Some previous work including (Ji and Grishman, 2006; Richman and Schone, 2008; Althobaiti et al., 2013) exploited semi-supervised methods to save annotation cost. We observed that self-training can provide further gains when the training data contains certain amount of noise.

6 Conclusions and Future Work

We developed a simple yet effective framework that can extract names from 282 languages and link them to an English KB. This framework follows a fully automatic training and testing pipeline, without the needs of any manual annotations or knowledge from native speakers. We evaluated our framework on both Wikipedia articles and external formal and informal texts and obtained promising results. To the best of our knowledge, our multilingual name tagging and linking framework is applied to the largest number of languages. We release the following resources for each of these 282 languages: "silver-standard" name tagging and linking annotations with multiple levels of granularity, morphology analyzer if it's a morphologically-rich language, and an end-to-end name tagging and linking system. In this work, we treat all languages independently when training their corresponding name taggers. 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.

Acknowledgments

This work was supported by the U.S. DARPA LORELEI Program No. HR0011-15-C-0115, ARL/ARO MURI W911NF-10-1-0533, DARPA DEFT No. FA8750-13-2-0041 and FA8750-13-2-0045, and NSF CAREER No. IIS-1523198. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. 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, pages 1024–1029. https://doi.org/10.3115/v1/N15-1107.

Fahd Alotaibi and Mark Lee. 2012. Mapping arabic wikipedia into the named entities taxonomy. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, pages 43–52. http://aclweb.org/anthology/C12-2005.

Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. 2013. 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, pages 32–40. http://aclweb.org/anthology/R13-1005.

Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. 2014. 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, pages 106–115. https://doi.org/10.3115/v1/E14-3012.

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. 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*. http://aclweb.org/anthology/P03-2031.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, pages 178–186. http://aclweb.org/anthology/W13-2322.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. ACM, New York, NY, USA, SIGMOD '08, pages 1247–1250. https://doi.org/10.1145/1376616.1376746.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*.

Seng Yee Chan and Dan Roth. 2010. Exploiting background knowledge for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, pages 152–160. http://aclweb.org/anthology/C10-1018.

Wanxiang Che, Mengqiu Wang, D. Christopher Manning, and Ting Liu. 2013. 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, pages 52–62. http://aclweb.org/anthology/N13-1006.

Wisam Dakka and Silviu Cucerzan. 2008. Augmenting wikipedia with named entity tags. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*. http://aclweb.org/anthology/I08-1071.

Luciano Del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. 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, pages 868–878. https://doi.org/10.18653/v1/D15-1103.

Asif Ekbal, Eva Sourjikova, Anette Frank, and Simone Paolo Ponzetto. 2010. Assessing the challenge of fine-grained named entity recognition and classification. In *Proceedings of the 2010 Named Entities Workshop*. Association for Computational Linguistics, pages 93–101. http://aclweb.org/anthology/W10-2415.

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *COLING 2002: The 19th International Conference on Computational Linguistics*. http://aclweb.org/anthology/C02-1130.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. Context-dependent fine-grained entity type tagging. *CoRR* abs/1412.1820. http://arxiv.org/abs/1412.1820.

Claudio Giuliano. 2009. 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, pages 201–209. http://aclweb.org/anthology/W09-1125.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. 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, pages 1177–1185. http://aclweb.org/anthology/C14-1111.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. 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, pages 110–120. https://doi.org/10.3115/v1/D14-1012.

Alexander Hogue, Joel Nothman, and James R. Curran. 2014. Unsupervised biographical event extraction using wikipedia traffic. In *Proceedings of the Australasian Language Technology Association Workshop 2014*. pages 41–49. http://aclweb.org/anthology/U14-1006.

Heng Ji and Ralph Grishman. 2006. Analysis and repair of name tagger errors. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics, pages 420–427. http://aclweb.org/anthology/P06-2055.

Heng Ji, Joel Nothman, and Hoa Trang Dang. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end kbp. In *Proceedings of the Text Analysis Conference*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics, Volume 29, Number 1, March 2003* http://aclweb.org/anthology/J03-1002.

Jun'ichi Kazama and Kentaro Torisawa. 2007. 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)*. http://aclweb.org/anthology/D07-1073.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. 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, pages 694–702. http://aclweb.org/anthology/P12-1073.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. 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, pages 260–270. https://doi.org/10.18653/v1/N16-1030.

Hao Li, Heng Ji, Hongbo Deng, and Jiawei Han. 2011. 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)*.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '12, pages 1727–1731. https://doi.org/10.1145/2396761.2398506.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'12, pages 94–100.

Patrick Littell, Kartik Goyal, R. David Mortensen, Alexa Little, Chris Dyer, and Lori Levin. 2016. 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, pages 998–1006. http://aclweb.org/anthology/C16-1095.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. Yago3: A knowledge base from multilingual wikipedias. In *Proceedings of the Conference on Innovative Data Systems Research*.

Alireza Mahmoudi, Mohsen Arabsorkhi, and Heshaam Faili. 2013. Supervised morphology generation using parallel corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. INCOMA Ltd. Shoumen, BULGARIA, pages 408–414. http://aclweb.org/anthology/R13-1053.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. 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, pages 55–60. https://doi.org/10.3115/v1/P14-5010.

James Mayfield, Dawn Lawrie, Paul McNamee, and Douglas W. Oard. 2011. 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*.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, pages 255–263. http://aclweb.org/anthology/I11-1029.

Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza, and Jordi Atserias. 2008. Learning to tag and tagging to learn: A case study on wikipedia. *IEEE Intelligent Systems* .

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. 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, pages 1003–1011. http://aclweb.org/anthology/P09-1113.

Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. 2013. 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, pages 1488–1497. http://aclweb.org/anthology/P13-1146.

Joel Nothman, R. James Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*. pages 124–132. http://aclweb.org/anthology/U08-1016.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence* 194:151–175. https://doi.org/10.1016/j.artint.2012.03.006.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1130–1139. https://doi.org/10.3115/v1/N15-1119.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. http://aclweb.org/anthology/N06-1025.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. 2015. 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*. ACM, New York, NY, USA, KDD '15, pages 995–1004. https://doi.org/10.1145/2783258.2783362.

E. Alexander Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, pages 1–9. http://aclweb.org/anthology/P08-1001.

Nicky Ringland, Joel Nothman, Tara Murphy, and R. James Curran. 2009. Classifying articles in english and german wikipedia. In *Proceedings of the Australasian Language Technology Association Workshop 2009*. pages 20–28. http://aclweb.org/anthology/U09-1004.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT, Short Papers*. Association for Computational Linguistics, pages 117–120. http://aclweb.org/anthology/P08-2030.

Teemu Ruokolainen, Oskar Kohonen, Kairit Sirts, Stig-Arne Grönroos, Mikko Kurimo, and Sami Virpioja. 2016. A comparative study of minimally supervised morphological segmentation. *Computational Linguistics* .

Avirup Sil and Radu Florian. 2016. 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, pages 2255–2264. https://doi.org/10.18653/v1/P16-1213.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 219–228. https://doi.org/10.18653/v1/K16-1022.

Mengqiu Wang, Wanxiang Che, and D. Christopher Manning. 2013. 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, pages 1073–1082. http://aclweb.org/anthology/P13-1106.

Mengqiu Wang and D. Christopher Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics* 2:55–66. http://aclweb.org/anthology/Q14-1005.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990* http://aclweb.org/anthology/J90-1003.

Dani Yogatama, Daniel Gillick, and Nevena Lazic. 2015. 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, pages 291–296. https://doi.org/10.3115/v1/P15-2048.

Amir Mohamed Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, pages 1361–1370. http://aclweb.org/anthology/C12-2133.

Boliang Zhang, Xiaoman Pan, Tianlu Wang, Ashish Vaswani, Heng Ji, Kevin Knight, and Daniel Marcu. 2016a. 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, pages 249–259. https://doi.org/10.18653/v1/N16-1029.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran XU. 2016b. 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, pages 461–470. http://aclweb.org/anthology/C16-1045.