# Embracing Non-Traditional Linguistic Resources for Low-resource Language Name Tagging

**Boliang Zhang**[1], **Di Lu**[1], **Xiaoman Pan**[1], **Ying Lin**[1],
**Halidanmu Abudukelimu**[2], **Heng Ji**[1], **Kevin Knight**[3]
[1] Computer Science Department, Rensselaer Polytechnic Institute
{zhangb8,lud2,panx2,liny9,jih}@rpi.edu
[2] Computer Science Department, Tsinghua University
abdklmhldm@gmail.com
[3] Information Sciences Institute, University of Southern California
knight@isi.edu

## Abstract

Current supervised name tagging approaches are inadequate for most low-resource languages due to the lack of annotated data and actionable linguistic knowledge. All supervised learning methods (including deep neural networks (DNN)) are sensitive to noise and thus they are not quite portable without massive clean annotations. We found that the F-scores of DNN-based name taggers drop rapidly (20%-30%) when we replace clean manual annotations with noisy annotations in the training data. We propose a new solution to incorporate many *non-traditional* language universal resources that are readily available but rarely explored in the Natural Language Processing (NLP) community, such as the World Atlas of Linguistic Structure, CIA names, PanLex and survival guides. We acquire and encode various types of non-traditional linguistic resources into a DNN name tagger. Experiments on three low-resource languages show that feeding linguistic knowledge can make DNN significantly more robust to noise, achieving 8%-22% absolute F-score gains on name tagging without using any human annotation [1].

## 1 Introduction

There is a general agreement that Deep Neural Networks provides a general, powerful underlying model for Information Extraction (IE), confirmed by improved state-of-the-art performance on many tasks such as name tagging (Chiu and Nichols, 2016; Lample et al., 2016), relation classification (Zeng et al., 2014; Liu et al., 2015; Nguyen and Grishman, 2015b; Yang et al., 2016) and event detection (Nguyen and Grishman, 2015b; Chen et al., 2015; Nguyen and Grishman, 2015a, 2016; Feng et al., 2016). For example, our experiments on several languages show that a DNN-based name tagger generally outperforms (up to 6% F-score gain) a Conditional Random Fields (CRFs) model trained from the same labeled data and feature set. DNN architecture is attractive to couple with character/word embeddings for IE tasks because it is easy to learn and usually effective enough to eliminate the need of explicit linguistic feature design.

However, training general models like DNN 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 (Zhang et al., 2016a). In order to compensate this data requirement, various automatic annotation generation methods have been proposed, including knowledge base driven distant supervision (An et al., 2003; Mintz et al., 2009; Ren et al., 2015), 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), and leveraging naturally existing noisy annotations such as Wikipedia markups (Nothman et al., 2008; Dakka and Cucerzan, 2008; Ringland et al., 2009; Alotaibi and Lee, 2012; Nothman et al., 2012; Althobaiti et al., 2014; Pan et al., 2017). Annotations produced from these methods are usually very noisy, while DNN is sensitive to noise just like many other machine learning methods. Our name tagging experiment shows that the F-score of the same DNN model learned from noisy training data is 20-30% lower than that trained from clean data. One major reason is that most of these methods solely rely on implicit embedding features in order to be (almost) language-independent.

Moreover, certain types of linguistic properties

[1] We make all cleaned resources and converted linguistic features publicly available at http://nlp.cs.rpi.edu/denoise

*Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 362–372,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

are difficult to be captured by embeddings, such as: (1) language-specific structures. For example, the Subject (S), Verb (V) and Object (O) orders in Tagalog are VS, VO, and VSO, which indicates that the word at the beginning of a sentence is usually a verb and thus unlikely to be a name. (2) culture-specific knowledge. For example, a Uyghur person's last name is the same as his/her father's first name.

On an almost parallel research avenue, linguists and domain experts have created a wide variety of multi-lingual resources, such as World Atlas of Linguistic Structure (WALS) (Dryer and Haspelmath, 2013b), Central Intelligence Agency (CIA) Names, grammar books, and survival guides. Such resources have been largely ignored by the mainstream statistical NLP research, because they were not specifically designed for NLP purpose at the first place and they are often far from complete. Thus they are not immediately actionable - converted into features, rules or patterns for a target NLP application. In this paper we design various methods to convert them into machine readable features for a new DNN architecture. Very little work has used non-traditional resources mentioned in this paper for practical downstream NLP applications. Limited work only used them for resource building (e.g., (Sarma et al., 2012)) or studying word order typology (Ostling, 2015). To the best of our knowledge, our work is the first to encode them as actionable knowledge for IE.

We aim to answer the following research questions: How to effectively acquire linguistic knowledge from non-traditional resources, and represent them for computational models? How much further gain can be obtained in addition to traditional resources?

## 2 Approach Overview

### 2.1 A Typical Baseline DNN Model

A typical supervised name tagger is presented in (Lample et al., 2016), consisted of Bi-directional Long Short-Term Memory networks (Bi-LSTM) and CRFs. We can consider name tagging as a sequence labeling problem, to tag each token in a sentence as the Beginning (B), Inside (I) or Outside (O) of a name mention with a certain type. In this paper we classify names into three types: person (PER), organization (ORG) and location (LOC). Predicting the tag for each token needs evidence from both of its previous context and future context in the entire sentence. Bi-LSTM networks (Graves et al., 2013) meet this need by processing each sequence in both directions with two separate hidden layers, which are then fed into the same output layer. Moreover, there are strong classification dependencies among name tags in a sequence. For example, "I-LOC" cannot follow "B-ORG". CRFs model, which is particularly good at jointly modeling tagging decisions, can be built on top of the Bi-LSTM networks.

| Languages | # of Documents | | # of Names | | # of Sentences | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| Hausa | 137 | 100 | 3,414 | 1,320 | 3,156 | 1,130 |
| Turkish | 128 | 100 | 2,341 | 2,173 | 1,973 | 2,119 |
| Uzbek | 127 | 100 | 3,577 | 3,137 | 3,588 | 3,037 |

Table 1: Data Statistics.

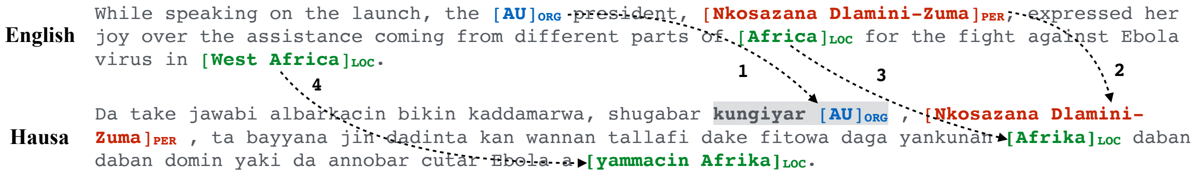
### 2.2 Baseline's Sensitiveness to Noise

In low-resource settings where few clean annotations are available, we could try to automatically generate some annotations to train the above model. For instance, we can project automatic annotations from a high-resource language (HL) to a low-resource language (LL) through parallel data. Figure 1 shows an example of projecting English automatic name annotations to Hausa through a parallel sentence pair.

We are interested in studying how sensitive DNN is to noise in such automatically generated training data. For our experiments we use English as the HL and use three LLs with different linguistic properties: Turkish, Uzbek and Hausa. We evaluate our approaches using the ground-truth name tagging annotations from the DARPA LORELEI program [2]. For fair comparison with previous LORELEI work (Tsai et al., 2016; Zhang et al., 2016a; Pan et al., 2017), we use the same 100 test documents. Table 1 shows detailed data statistics.

We use 80% of the name annotated LL documents for training and 20% for development, and parallel sentences to artificially create noisy training data as follows. We use $S$ to denote the sentences in LL and $T$ to denote the sentences in HL. We apply Stanford English name tagger (Manning et al., 2014) on $T$ and project English names onto $S$, using the following measurements to determine whether a candidate LL name string $n_l$ matches an expected English name $n_e$: (1) If the edit distance

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


English: While speaking on the launch, the [AU]$_{\text{ORG}}$ president, [Nkosazana Dlamini-Zuma]$_{\text{PER}}$, expressed her joy over the assistance coming from different parts of [Africa]$_{\text{LOC}}$ for the fight against Ebola virus in [West Africa]$_{\text{LOC}}$.

Hausa: Da take jawabi albarkacin bikin kaddamarwa, shugabar kungiyar [AU]$_{\text{ORG}}$ , [Nkosazana Dlamini-Zuma]$_{\text{PER}}$ , ta bayyana jin dadinta kan wannan tallafi dake fitowa daga yankunan [Afrika]$_{\text{LOC}}$ daban daban domin yaki da annobar cutar Ebola a [yammacin Afrika]$_{\text{LOC}}$.

* Projection 1 is incorrect and results in a noisy instance in the automatically generated Hausa annotations. The correct name mention is "kungiyar AU (Africa Union)" instead of "AU".

Figure 1: Noisy Training Data Generation by Projecting English Automatic Name Annotations to Hausa.

between $n_e$ and $n_l$ is not greater than two. (2) We check the pronunciations of $n_e$ and $n_l$ based on Soundex (Odell, 1956), Metaphone (Philips, 1990) and NYSIIS (Taft, 1970) algorithms. We consider two codes match if their edit distance is not greater than two. (3) If $n_e$ and $n_l$ are aligned in the parallel data by running GIZA++ word alignment tool (Och and Ney, 2003).

In this way we obtain an automatically generated noisy training data set $Train_{noise}$. We denote $Train_{clean}$ as the ground truth which is manually created by human annotators on set $S$. We mix $Train_{noise}$ and $Train_{clean}$ in different proportions to obtain a training set $Train_{mix}$ on various noise levels. We define ***noise level*** as $1 - fscore(Train_{mix})$ where the f-score of $Train_{mix}$ is computed against $Train_{clean}$. For example, when $Train_{mix}$ is full of manually created clean data, the noise level is 0; when we mix half $Train_{noise}$ and half $Train_{clean}$ of the Hausa data, the f-score of $Train_{mix}$ is 80.1%, and the noise level is 19.9%.

To learn embeddings, we use 12,624 Hausa documents from the LORELEI program, and use 288,444 Turkish documents and 128,763 Uzbek documents from a June 2015 Wikipedia dump. Figure 2 shows the performance of the baseline tagger trained from $Train_{mix}$ for three languages. We can clearly see that the performance drops rapidly as the training data includes more noise.

### 2.3 A New Improved Model

We propose to acquire non-traditional linguistic resources and encode them as new actionable features (Section 3). In Figure 3, we design three integration methods to incorporate explicit linguistic features into Bi-LSTM networks: (1) concatenate the linguistic features and word embeddings at the input level, (2) concatenate the linguistic features and the bidirectional encodings of each token before feeding them into the output layer that computes the tag probability, and (3) use an additional Bi-LSTM to consume the feature embeddings of each token and concatenate both Bi-LSTM encodings of feature embeddings and word embeddings before the output layer. We set the word input dimension to 100, word LSTM hidden layer dimension to 100, character input dimension to 50, character LSTM hidden layer dimension to 25, input dropout rate to 0.5, and use stochastic gradient descent with learning rate 0.01 for optimization.

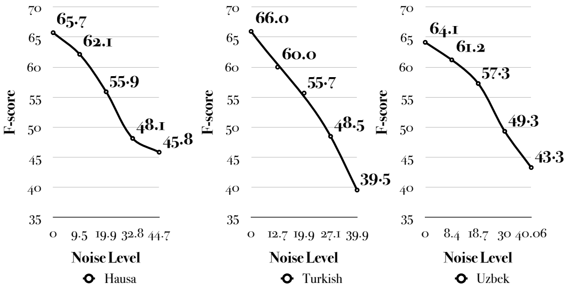


Figure 2: Performance of baseline DNN Name Taggers Trained from Data with Various Noise Levels (The noise level is created by assigning the proportion of $Train_{noise}$ in $Train_{mix}$ as 0%, 25%, 50%, 75% and 100% respectively. )

## 3 Incorporating Non-traditional Linguistic Knowledge

In this section we will describe the detailed methods to acquire and encode various types of non-traditional resources. We call them as *non-traditional* because they have been rarely used in previous NLP research.

### 3.1 Basic Knowledge about the Language

**Wikipedia Description.** An English Wikipedia page about a language usually provides us general descriptions of the language. In particular, the list of usable characters, gender indicators, capitalization information, transliteration and number spelling rules are most useful for name tagging. The list of usable characters for regular words in a particular language can help us detect foreign borrow words, which are likely to be names. For example, "*th*" usually does not appear at the begin-

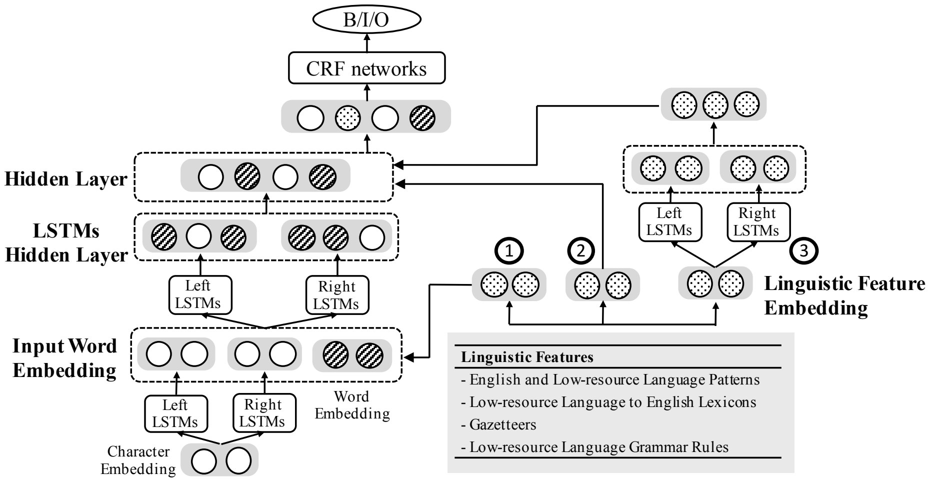


Figure 3: Three Integration Methods to Incorporate Explicit Linguistic Features into DNN.

ning of a Turkish word. Thus "*Thomas Marek*" is likely to be a foreign name.

**Grammar Book.** From grammar books we can also extract more language-specific contextual words, prefixes, suffixes and stemming rules. Name related lists contain: case suffix, preposition, postposition, ordinal number, definite article, negation, conjunction, pronoun, quantifier, numeral, time, locative, question particle, demonstrative, degree word, plural prefix/suffix, subordinator, reduplication, possessive, situational and epistemic markers. Table 2 shows some examples of name related suffix features.

### 3.2 Linguistic Structure

Recently linguists have made great efforts at building linguistic knowledge bases for thousands of languages in the world. Two such examples are WALS database (Dryer and Haspelmath, 2013a) and Syntactic Structures of the World's Languages [3]. These databases classify languages according to a large number of topological properties (phonological, lexical and grammatical). For example, WALS consists of 141 maps with accompanying text on diverse properties, gathered from descriptive materials (such as reference grammars). Altogether there are 2,676 languages and more than 58,000 data points; each data point is a (language, feature, feature value) tuple that specifies the value of the feature in a particular language. (e.g., (English, canonical word order, SVO)). In total we extract 188 linguistic properties related to name tagging, belonging to 20 Phonology, 13 Lexicon, 12 Morphology, 29 Nominal, 8 Nominal Syntax, 17 Verbal Categories, 56 Word Order, 26 Simple Clauses, and 7 Complex Sentences categories respectively. Table 3 shows some examples.

### 3.3 Multi-lingual Dictionaries

**CIA Names.** We utilize the CIA Name Files [4], which include biographical sketches, memorandums, telegrams, legislative records, legal documents, statements, and other records. We used the version cleaned up by Lawson et al. [5] that includes documents about names in 41 languages. Besides, person names in certain regions often include some common syllable patterns. Table 4 presents some examples. In languages such as Turkish, Uzbek and Uyghur, a person's last name inherits from his or her father's first name. In Uyghur, there are no additional suffixes. In Uzbek, additional suffixes include "*-ov*", "*-ev*", "*-yev*", "*-eva*" and "*-yeva*". In Turkish, a male's first name often ends with a consonant, and his last name consists of his father's first name and a suffix "*-oğlu* (son of)". We exploit this kind of knowledge to improve gazetteer match and name boundary identification.

**Unicode CLDR.** Unicode Common Locale Data Repository (CLDR) [6] is a data collection for 194 languages, maintained by the Unicode Consortium to support software internationalization and localization. We extract bi-lingual location gazetteers, and exploit patterns and lists of currencies, months, weekdays, day periods and time units to remove them from name candidates because they share some features with names (e.g., capitalization, "*Ocak*" in Turkish means "*January*").

[3] http://sswl.railsplayground.net/

[4] https://www.archives.gov/iwg/declassified-records/rg-263-cia-records

[5] https://www.researchgate.net/profile/Edwin_Lawson

[6] http://cldr.unicode.org/

| Languages | Features | Description | Examples |
|---|---|---|---|
| Uzbek | Name | **-ni** (accusative), **-ning** (possessive), **-da** (locative), **-dan** (ablative) | **Turkiyaning** (of Turkey), **Turkiyada** (in Turkey), **Turkiyaga** (to Turkey), **Turkiyadan** (from Turkey). |
| | Non-Name | Suffix **-roq** indicates adjectives<br>Suffixes **-lar/-ler** indicate plurals | **qoraroq** (darker)<br>**qizlar** (daughters) |
| Hungarian | Name | Foreign name with >1 tokens and an adjective marker | New York-**i** (from New York) |
| | | Most names with adjective or verbal suffix are lowercased | Balzac + **-os** ⇒ balzac**os** |
| | | Possession relation | Péter-**ék** (Peter and his group), Péter-**é** (that of Peter) |
| | | Affixes associated with names | Sartre-**nak** (to Sartre), Bordeaux-**ban** (in Bordeaux), Smith-**ért** (for Smith) |
| | Non-Name | Non-Name POS tag<br>Complete inflectional for nominals | adjectives (**-tlen**: "**-less**"), verbs tense (**meg-**:"completed"), conjunctions (**-ért**: "because of")<br>karoknak (for arms) → karok (arms) → kar (arm) |
| Uyghur | Name | Animacy suffixes | **ning**, **ni**, **luq**, and **lik** |
| | | Geopolitical or location suffixes | **ke**, **ge**, **qa**, **gha**, **te**, **de**, **ta**, **da**, **tin**, **din**, **tiki**, **diki**, **kiche**, **giche**, **qiche**, and **ghiche**. |
| Turkish | Name | Postpositions | karaköy**de** (in Karaköy) |

Table 2: Name-related Knowledge Summarized from Grammar Books.

| Languages | Categories | Description | Name Related Characteristics |
|---|---|---|---|
| Tagalog | Subject, Verb, Object Order | VS, VO, VSO | the word at the beginning of a sentence is unlikely to be a name |
| Turkish | Negation | Suffix -me at the root of a verb indicates negations | not a name |
| Bengali | Animacy | -ta is a case that indicates inanimacy | |
| Thai | Nested Name Structure | Delimiter between modifier and head, [ORG กระทรวงต่างประเทศ] **ของ**[LOC อินโดนีเซีย] ([ORG Foreign Ministry ] of [LOC Indonesia]) | Name boundary |
| Tamil | Conjunction Structure | Name1-**yum** Name2-**yum** (Name1 and Name2) | Name type consistency |

Table 3: Name-related Knowledge Extracted from WALS.

| Languages | Frequent Syllable Patterns | Examples |
|---|---|---|
| Slavic | Suffixes: -ov, -ev -ova, -eva; -ovich, -ich, -enko, -ko, -chuk, -yuk, -ak, -chenko, -skiy, -ski, -vych, -vich | Karimov, Yuriy Yarov, Abdulaziz Komilov, Yamonkulov Yaxshiboyevich, Shevchenko |
| Arabic | Prefixes: al-, Ahl, Abdul-, Abdu-<br>Suffixes: -allah, -ullah | Abdul Khaliq, Abdul Latif, Abdul Maajid<br>Daifallah, Dhikrullah, Faizullah, Fathallah |
| Uzbek | Suffixes: -ov, -ova, -ev -yev, -eva, -yeva; -ovich, -evich, -ich | Karim Ahmedov, Ahmed Aliev, Zulfiya Karimova, Karmm Sharafovich Rashidov |

Table 4: Common Syllable Patterns Extracted from CIA Names.

**Wiktionary.** Wiktionary [7] is a web-based collaborative project to create an English content dictionary of all words in many languages. We collected dictionaries in 1,247 languages.

**Panlex.** Panlex [8] (Baldwin et al., 2010; Kamholz et al., 2014) database contains 1.1 billion pairwise translations among 21 million expressions in about 10,000 language varieties.

**Multilingual WordNet.** We leverage three versions of multi-lingual WordNet: (1) Open Multilingual WordNet (Bond and Paik, 2012) which links words in many languages to English WordNet based on Wiktionary and CLDR; (2) Universal WordNet (de Melo and Weikum, 2019) which automatically extends English WordNet with around 1.5 million meaning links for 800,000 words in over 200 languages, based on WordNets, translation dictionaries and parallel corpora; and (3) Etymological WordNet (de Melo and Weikum, 2010; de Melo, 2014) that provides information about how words in various languages are etymologically related based on Wiktionary.

**Phrase Pairs Mined from Wikipedia.** From Wikipedia we extracted all pairs of titles that are connected by cross-lingual links. And we extracted more phrase translation pairs using parenthesis patterns from the beginning sentences of Wikipedia pages. For example, from the first sentence of the English Wikipedia page about Ürümqi: "*Ürümqi (ئۈرۈمچى) is the capital of the*

[7] https://en.wiktionary.org

[8] http://panlex.org/

*Xinjiang Uyghur Autonomous Region of the People's Republic of China in Northwest China,*" we can extract an Uyghur-English name translation pair of "ئۈرۈمچى" and "*Ürümqi*". Moreover, we retrieved related Wikipedia articles, and mined common names in many languages and regions.

**GeoNames.** We exploit the geo-political and location entities in multilingual GeoNames database [9]. It contains over 10 million geographical names and over 9 million unique features of the following properties: id, name, asciiname, alternate names, latitude, longitude, feature class, feature code, country code, administrative code, population, elevation and time zone.

**JRC Names.** Finally we include the JRC Names (Steinberger et al., 20011), a large list of person and organization names (about 205,000 entries) in over 20 different scripts. Some entries include additional information such as frequency, title and date ranges.

**Grounding to KB and Typing.** For names that we are able to acquire English translations, we further ground ("wikify") them to an external knowledge base (KB, DBpedia in our work) if they are linkable. We use two measures (Pan et al., 2015) for linking: (1) Popularity: we prefer popular entities in the KB; (2) Coherence: we link a pair of a foreign name and its English translation simultaneously and favor their candidate entities that are also strongly connected in the KB through a direct cross-lingual page link, a common neighbor, or sharing similar properties. After linking, we assign an entity type to each pair based on their properties in the KB (e.g., an entity with a birth-date and a death-date is likely to be a person). The typing component is a Maximum Entropy model learned from the Abstract Meaning Representation (Banarescu et al., 2013) corpus that includes both entity type and Wikipedia link for each entity mention, using KB properties as features.

### 3.4 Phrase Books

Finally we exploit phrase books that include phrase translations between many languages and English.

**Language Survival Kits.** FAMiliarization [10] offers language survival kits (LSKs) for 100 languages, each of which has up to 10 kits of different topics. LSK encodes phrases, translations, and romanizations and is available for 55 languages. FAMiliarization also provides translations of name-related words and phrases.

| Language | Gazetteer | | | Title | Non-Name | Suffix |
|---|---|---|---|---|---|---|
| | PER | LOC | ORG | | | |
| Hausa | 1,174 | 5,123 | 199 | 42 | 391 | 21 |
| Turkish | 2,819 | 7,271 | 262 | 231 | 411 | 181 |
| Uzbek | 1,771 | 5,331 | 103 | 178 | 271 | 209 |

Table 5: Name Related List Statistics (# of entries).

For each language, we first extracted $2,000$ to $3,000$ parallel sentence/phrase pairs. Then we ran GIZA++ over these pairs and combined structure rules from WALS to obtain word translation pairs. We also extracted translations of the following English lists: cardinal number, currency, disease, location affixes, title, nationalities, topical keywords, organization suffixes, temporal words, locations and people, and stop words which are unlikely to be names.

**Elicitation Corpus.** An elicitation corpus is a controlled corpus translated by a bilingual consultant in order to produce high quality word aligned sentence pairs. During the elicitation process, the user will translate a subset of these sentences that is dynamically determined to be sufficient for learning the desired grammar rules. We extracted word and phrase translation pairs from the Elicitation corpus developed by CMU (Probst et al., 2001; Alvarez et al., 2005) [11] for the DARPA LORELEI which contains pairs of sentences in a low-resource language and English.

### 3.5 Encoding Linguistic Features

We merged the linguistic resources collected above into three types of features: (1) name gazetteers; (2) list of suffixes and contextual words (e.g., titles) that indicate names; and (3) list of words that indicate non-names (e.g., time expressions). Ultimately we obtained 30 explicit linguistic feature categories. Table 5 shows the statistics of the encoded features.

For each token $w_i$ in a sentence, we check whether $w_i$, its previous token $w_{i-1}$ and its next token $w_{i+1}$ exist in these lists, and concatenate them into an initial feature vector for $w_i$. For any resources (e.g., lexicons and phrase books) that contain English translations, we also use them to translate each $w_i$, and check whether its translation is capitalized or exists in English name tagging resources (contextual words, gazetteers), whether its contexts match any English patterns as described

[9] http://www.geonames.org/

[10] http://fieldsupport.dliflc.edu/

[11] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/cmt-40/Nice/Elicitation/Elicitation_Corpus-LDC/

in (Zhang et al., 2016a).

## 4 Experiments

Using the data sets mentioned in Section 2.2, we conduct experiments for three languages: Hausa, Turkish and Uzbek.

### 4.1 Overall Performance

Table 6 compares the results of three feature integration methods described in Section 2.3 and Figure 3. We can see that the third integration method (Integration 3) consistently outperforms the others for all three languages.

| Models | Hausa | Turkish | Uzbek |
|---|---|---|---|
| Bi-LSTMs | 65.7 | 65.9 | 64.1 |
| + Integration 1 | 71.1 | 71.8 | 67.4 |
| + Integration 2 | 71.5 | 73.1 | 67.2 |
| + Integration 3 | **72.2** | **74.3** | **68.4** |

Table 6: Feature Integration Methods Comparison.

We compare the following models: a baseline model that uses only character and word embedding features, a model adding traditional linguistic features as described in (Zhang et al., 2016a), and a model further adding non-traditional linguistic features using the third integration method. Figure 4 presents the results. Clearly models trained with linguistic features substantially outperform the baseline models on all noise levels for all languages. As the noise level increases, the performance of the baseline model drops drastically while the model trained with linguistic features successfully curbs the downward trend and forms a relatively flat curve at last. Adding non-traditional linguistic features provides further gains in almost all settings. Notably for Turkish, adding linguistic features and using 100% automatically generated noisy training data, our approach achieves the same performance as the baseline model using 75% manually created clean data and 25% automatically created noisy data. In other words, explicit linguistic knowledge has significantly saved annotation cost (2,367 sentences). Our results without using any manually labeled training data are much better than state-of-the-art reported in our previous work (Zhang et al., 2016a) which used most traditional resources mentioned in this paper and (Pan et al., 2017) which derived noisy training data from Wikipedia markups. On the same test sets we achieved 5.5% higher F-score for Hausa than (Zhang et al., 2016a), 27.7% higher F-score for Turkish and 13.6% higher F-score for Uzbek than (Pan et al., 2017).

| | Category | Hausa | Turkish | Uzbek |
|---|---|---|---|---|
| **A** | Embedding feature | 45.8 | 39.5 | 43.3 |
| **B** | (A)+Pattern mining and projection | 46.7 | 40.9 | 45.4 |
| **C** | (B)+Basic knowledge and linguistic structure | 50.4 | 53.3 | 52.4 |
| **D** | (C)+Dictionaries | 52.0 | 57.7 | 56.1 |
| **E** | (D)+Phrase books | 53.8 | 60.0 | 57.8 |

Table 7: Contributions of Various Categories of Linguistic Knowledge (F-score (%)).

### 4.2 Detailed Analysis

Table 7 presents the contribution of each linguistic feature category when using 100% automatically created training data. Figure 5 shows some examples of errors corrected by each category. Some remaining challenges pertain to the lack of contextual clues for identifying the boundaries of long organizations, especially when they include nested or conjunction structures (e.g., "*Uluslararası ve Stratejik Araştırmalar Merkezi'nde (International and Strategic Research Center)*" in Turkish). The performance of organization tagging is 16%-31% lower than that of persons and locations. We also observe a "*popularity bias*" challenge, especially because we don't have enough resources and tools to perform a deep understanding of the contexts. For example, when a journal name "*New England*" appears in Hausa texts, all of its mentions are mistakenly labeled as location instead of organization, because the dominant type label of "*New England*" is location in all of our resources.

## 5 Related Work

The major novel contribution of this paper is to systematically explore many non-traditional linguistic resources which have been largely neglected by the mainstream NLP community. Some previous efforts used WALS to study the typological relations across languages (Rama and Prasanth, 2012; O'Horan et al., 2016; Yamauchi and Murawaki, 2016) but very little work used it for practical NLP applications. Most DNN methods solely relied on character embeddings and word embeddings as features for name tagging (e.g., (Huang et al., 2015; Lample et al., 2016; Chiu and Nichols, 2016)). (Shimaoka et al., 2017) used hand-crafted features to improve the performance of DNN on fine-grained entity typing. (Chiu and Nichols, 2016) attempted to incorporate gazetteers as ex-

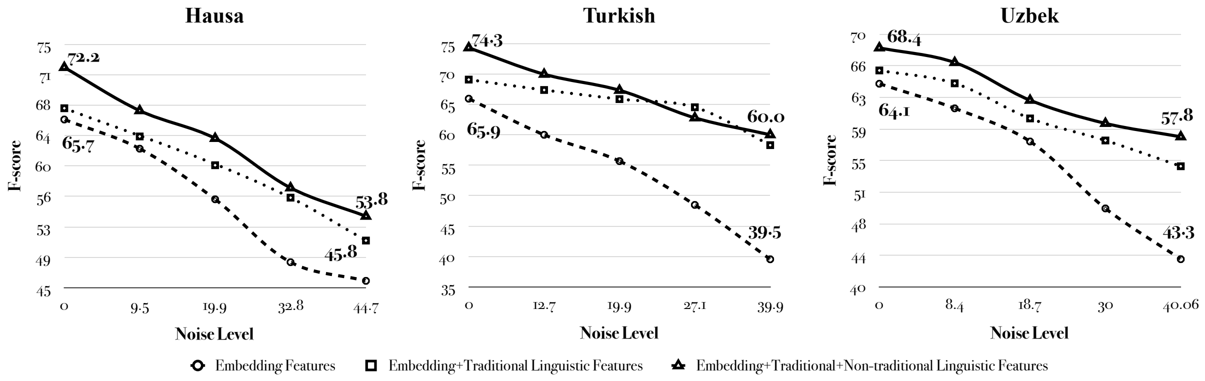
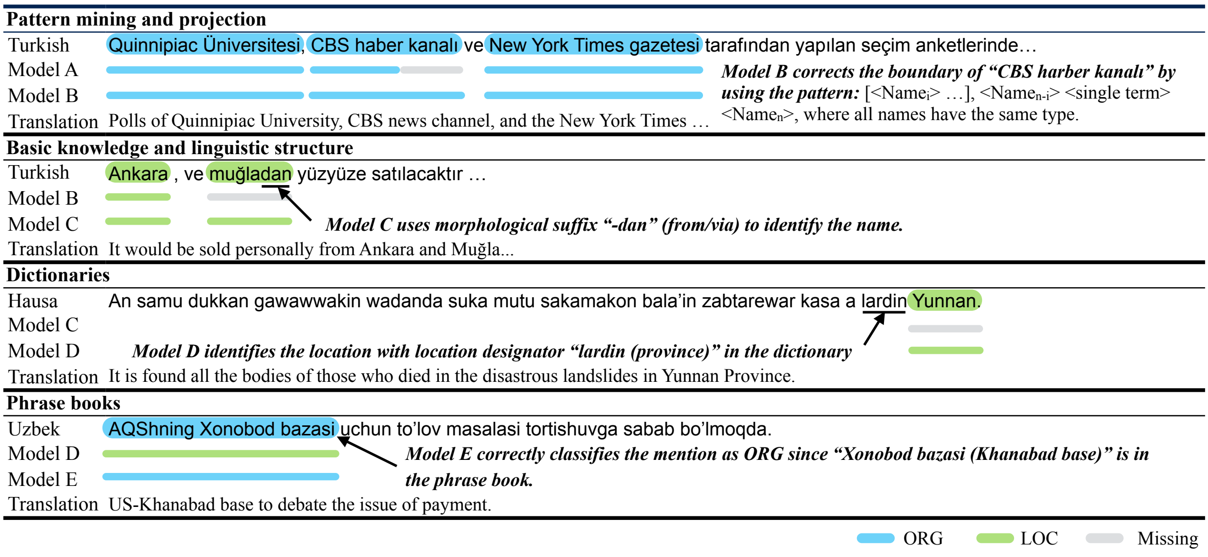


Figure 4: Name Tagging Performance.



**Pattern mining and projection**

Turkish Quinnipiac Üniversitesi, CBS haber kanalı ve New York Times gazetesi tarafından yapılan seçim anketlerinde…

Model A

Model B

*Model B corrects the boundary of "CBS harber kanalı" by using the pattern:* [<Name$_i$> …], <Name$_{n-i}$> <single term> <Name$_n$>, where all names have the same type.

Translation Polls of Quinnipiac University, CBS news channel, and the New York Times …

**Basic knowledge and linguistic structure**

Turkish Ankara , ve muğladan yüzyüze satılacaktır …

Model B

Model C

*Model C uses morphological suffix "-dan" (from/via) to identify the name.*

Translation It would be sold personally from Ankara and Muğla...

**Dictionaries**

Hausa An samu dukkan gawawwakin wadanda suka mutu sakamakon bala'in zabtarewar kasa a lardin Yunnan.

Model C

Model D

*Model D identifies the location with location designator "lardin (province)" in the dictionary*

Translation It is found all the bodies of those who died in the disastrous landslides in Yunnan Province.

**Phrase books**

Uzbek AQShning Xonobod bazasi uchun to'lov masalasi tortishuvga sabab bo'lmoqda.

Model D

Model E

*Model E correctly classifies the mention as ORG since "Xonobod bazasi (Khanabad base)" is in the phrase book.*

Translation US-Khanabad base to debate the issue of payment.

Figure 5: Examples of Corrections Made by Each Category of Linguistic Knowledge.

plicit linguistic features, and found that gazetteers are not very effective when they have a low coverage of name variants or when they contain many ambiguous entries. We addressed this challenge by integrating gazetteers gathered from a much wider range of sources.

Some recent studies (Zhang et al., 2016a; Littell et al., 2016a; Tsai et al., 2016; Pan et al., 2017) under the DARPA LORELEI program focused on name tagging for low-resource languages. Most noise tolerant supervised learning algorithms (Bylander, 1994; Dredze et al., 2008; Crammer et al., 2009; Kalapanidas et al., 2003; Scott et al., 2013) have been applied for improving image classification (Mnih and Hinton, 2012; Natarajan et al., 2013; Sukhbaatar et al., 2014; Xiao et al., 2015). Coupling our idea with these algorithms is also likely to yield further improvement.

## 6 Conclusions and Future Work

Using name tagging as a case study, we demonstrated the power of acquiring and encoding non-traditional linguistic resources. Experiments showed that they can significantly improve the quality of supervised models like DNNs and make them much more robust to noise in automatically created training data. Recent trend of DNN research in the NLP community boasts getting rid of explicit feature design. Our work argues that data-driven implicit knowledge like word embeddings cannot cover all linguistic phenomena in low-resource settings. We propose to embrace the readily available universal resources for many languages, and proved this process of making them actionable is not costly and does not require a system developer to "know" the language. Many more non-traditional linguistic resources remain to explore in the future, including Lexvo (de Melo, 2015), Multilingual Entity Taxonomy (de Melo and Weikum, 2010), EZGlot, URIEL knowledge base (Littell et al., 2016b), travel phrase books and yellow phone books. We will also investigate whether these linguistic resources can make DNN more robust to other factors such as data size and topical relatedness.

## Acknowledgments

This work was supported by the U.S. DARPA LORELEI Program No. HR0011-15-C-0115. 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

Fahd Alotaibi and Mark Lee. 2012. Mapping arabic wikipedia into the named entities taxonomy. In *Proceedings of the International Conference on Computational Linguistics*.

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*.

Alison Alvarez, Lori Levin, Robert Frederking, Jeff Good, and Erik Peterson. 2005. Semi-automated elicitation corpus generation. In *Proceedings of MT Summit X*.

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*.

Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. Panlex and lextract: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

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 *ACL Workshop on Linguistic Annotation and Interoperability with Discourse*.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference*.

Tom Bylander. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*.

Wanxiang Che, Mengqiu Wang, Christopher D 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*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multipooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. In *Transaction of Association for Computational Linguistics*.

Koby Crammer, Alex Kulesza, and Mark Dredze. 2009. Adaptive regularization of weight vectors. In *Advances in neural information processing systems*.

Wisam Dakka and Silviu Cucerzan. 2008. Augmenting wikipedia with named entity tags. In *Proceedings of the International Joint Conference on Natural Language Processing*.

Gerard de Melo. 2014. Etymological wordnet: Tracing the history of words. In *Proceeddings of the Conference on Language Resources*.

Gerard de Melo. 2015. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web* 6:4.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In *Proceedings of the 5th Global Wordnet Conference*.

Gerard de Melo and Gerhard Weikum. 2019. Towards a universal wordnet by learning from combined evidence. In *Proceeddings of The Conference on Information and Knowledge Management*.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *Proceedings of the 25th international conference on Machine learning*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013a. *WALS Online*.

Matthew S. Dryer and Martin Haspelmath. 2013b. The world atlas of language structures online. In *Leipzig: Max Planck Institute for Evolutionary Anthropology*.

Xiaocheng Feng, Heng Ji, Duyu Tang, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceeddings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Alan Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding, 2013 IEEE Workshop on*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

Elias Kalapanidas, Nikolaos Avouris, Marian Craciun, and Daniel Neagu. 2003. Machine learning algorithms: A study on noise sensitivity. In *Proceeddings of 1st Balcan Conference in Informatics*.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

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*.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceeddings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*.

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 Conference on Information and Knowledge Management*.

Patrick Littell, Kartik Goyal, David Mortensen, Alexa Little, Chris Dyer, and Lori Levin. 2016a. Named entity recognition for linguistic rapid response in low-resource languages: Sorani kurdish and tajik. In *Proceedings of the Conference on Computational Linguistics*.

Patrick Littell, David Mortensen, and Lori Levin (eds.). 2016b. Uriel typological database. *Pittsburgh: Carnegie Mellon University (Available online at http://www.cs.cmu.edu/ dmortens/uriel.html)* .

Yang Liu, Furu Wei, Sujian Li, Heng Ji, and Ming Zhou. 2015. A dependency-based neural network for relation classification. In *Proceeddings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceeddings of the conference of the Association for Computational Linguistics*.

Volodymyr Mnih and Geoffrey E Hinton. 2012. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning*.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*.

Thien Huu Nguyen and Ralph Grishman. 2015a. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Thien Huu Nguyen and Ralph Grishman. 2015b. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of NAACL Workshop on Vector Space Modeling for NLP*.

Thien Huu Nguyen and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* .

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* .

Margaret King Odell. 1956. *The Profit in Records Management*. Systems (New York).

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of the International Conference on Computational Linguistics*.

Robert Ostling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

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*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics*.

Lawrence Philips. 1990. Hanging on the metaphone. *Computer Language* 7(12).

Katharina Probst, Ralf D. Brown, Jaime G. Carbonell, Alon Lavie, and Lori Levin. 2001. Design and implementation of controlled elicitation for machine translation of low-density languages. In *Proceedings of Workshop MT2010 at Machine Translation Summit VIII*.

Taraka Rama and Kolachina Prasanth. 2012. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of the International Conference on Computational Linguistics*.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, Heng Ji, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceeddings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Classifying articles in english and german wikipedia. In *Proceedings of Australasian Language Technology Association Workshop 2009*.

Shikhar Kr. Sarma, Dibyajyoti Sarmah, Biswajit Brahma, Mayashree Mahanta, Himadri Bharali, and Utpal Saikia. 2012. Building multilingual lexical resources using wordnets: Structure, design and implementation. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*.

Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *the Conference On Learning Theory*.

Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2017. Neural architectures for fine-grained entity type classfication. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, and Erik van der Goot. 20011. Jrc-names: A freely available, highly multilingual named entity resource. In *Proceeddings of the 8th International Conference on Recent Advances in Natural Language Processing*.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080* .

Robert L Taft. 1970. *Name Search Techniques*. New York State Identification and Intelligence System, Albany, New York, US.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the Conference on Natural Language Learning*.

Mengqiu Wang, Wanxiang Che, and Christopher D Manning. 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the Association for Computational Linguistics*.

Mengqiu Wang and Christopher Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. In *Transactions of the Association of Computational Linguistics*.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kenji Yamauchi and Yugo Murawaki. 2016. Contrasting vertical and horizontal transmission of typological features. In *Proceedings of the International Conference on Computational Linguistics*.

Yunlun Yang, Yunhai Tong, Shulei Ma, and Zhi-Hong Deng. 2016. A position encoding convolutional neural network based on dependency tree for relation classification. In *Proceedings of the Empirical Methods on Natural Language Processing*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceeddings of the 25th International Conference on Computational Linguistics*.

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 *Proceeddings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, and Heng Ji. 2016b. Bitext name tagging for annotation projection. In *Proceedings of the 26th International Conference on Computational Linguistics*.