Platforms for Non-Speakers Annotating Names in Any Language

Ying Lin,[1] **Cash Costello,**[2] **Boliang Zhang,**[1] **Di Lu,**[1]
Heng Ji,[1] **James Mayfield,**[2] **Paul McNamee**[2]
[1] Rensselaer Polytechnic Institute
{liny9,zhangb8,lud2,jih}@rpi.edu
[2] Johns Hopkins University
{ccostel2,mayfield,mcnamee}@jhu.edu

Abstract

We demonstrate two annotation platforms that allow an English speaker to annotate names for any language without knowing the language. These platforms provided high-quality "*silver standard*" annotations for low-resource language name taggers (Zhang et al., 2017) that achieved state-of-the-art performance on two surprise languages (Oromo and Tigrinya) at LoreHLT2017[1] and ten languages at TAC-KBP EDL2017 (Ji et al., 2017). We discuss strengths and limitations and compare other methods of creating silver- and gold-standard annotations using native speakers. We will make our tools publicly available for research use.

1 Introduction

Although researchers have been working on unsupervised and semi-supervised approaches to alleviate the demand for training data, most state-of-the-art models for name tagging, especially neural network-based models (Pan et al., 2017; Zhang et al., 2017) still rely on a large amount of training data to achieve good performance. When applied to low-resource languages, these models suffer from data sparsity. Traditionally, native speakers of a language have been asked to annotate a corpus in that language. This approach is uneconomical for several reasons. First, for some languages with extremely low resources, it's not easy to access native speakers for annotation. For example, Chechen is only spoken by 1.4 million people and Rejiang is spoken by 200,000 people. Second, it is costly in both time and money to write an annotation guideline for a low-resource language and to train native speakers (who are usually not linguists) to learn the guidelines and qualify for annotation tasks. Third, we observed poor annotation quality and low inter-annotator agreement among newly trained native speakers in spite of high language proficiency. For example, under DARPA LORELEI,[2] the performance of two native Uighur speakers on name tagging was only 69% and 73% F_1-score respectively.

Previous efforts to generate "silver-standard" annotations used Web search (An et al., 2003), parallel data(Wang and Manning, 2014), Wikipedia markups (Nothman et al., 2013; Tsai et al., 2016; Pan et al., 2017), and crowdsourcing (Finin et al., 2010). Annotations produced by these methods are usually noisy and specific to a particular writing style (e.g., Wikipedia articles), yielding unsatisfactory results and poor portability.

It is even more expensive to teach English-speaking annotators new languages. But can we annotate names in a language we don't know? Let's examine a Somali sentence:

> "*Sida uu saxaafadda u sheegay Dr* ***Jaamac Warsame Cali*** *oo fadhigiisu yahay magaalada* ***Baardheere*** *hadda waxaa shuban caloolaha la yaalla xarumaha caafimaadka 15-cunug oo lagu arkay fuuq bax joogto ah, wuxuu xusay dhakhtarku in ay wadaan dadaallo ay wax kaga qabanayaan xaaladdan*"

Without knowing anything about Somali, an English speaker can guess that "*Jaamac Warsame Cali*" is a person name because it's capitalized, the

We thank Kevin Blissett and Tongtao Zhang from RPI for their contributions to the annotations used for the experiments. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contracts No. HR0011-15-C-0115 and No. HR0011-16-C-0102. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

[1] https://www.nist.gov/itl/iad/mig/lorehlt-evaluations

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

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 1–6
Melbourne, Australia, July 15 - 20, 2018.

word on its left, "*Dr*," is similar to "*Dr.*" in English, and its spelling looks similar to the English "*Jamac Warsame Ali*." Similarly, we can identify "*Baardheere*" as a location name if we know that "*magaalada*" in English is "*town*" from a common word dictionary, and its spelling is similar to the English name "*Bardhere*."

What about languages that are not written in Roman (Latin) script? Fortunately language universal romanization (Hermjakob et al., 2018) or transliteration[3] tools are available for most living languages. For example, the following is a Tigrinya sentence and its romanized form:

> "ናይዚ እዋን'ዚ ፕረዝደንት ዑበደልፈ.ታሕ አል-ሲሲ ነቲ ናይ 2011 ዓ.ም.ፈ. ተቃውሞ ብምንኣድ እቲ ተቃውሚ ሓዳስ ግብጺ ዘምጸእ'ዩ ኢሎም።"
>
> "*naayezi 'ewaane'zi perazedanete* ***'aabedale-fataahhe 'ale-sisi*** *nati naaye 2011 'aa.me.fa taqaawemo bemene'aade 'eti taqaawemi hhaadaase gebetsi zametsa'a 'yulome* ."

An English speaker can guess that "**ዑበደልፈ.ታሕ አል-ሲሲ**" is a person name because its romanized form "*aabedalefataahhe 'ale-sisi*" sounds similar to the English name "*Abdel-Fattah el-Sissi*," and the romanized form of the word on its left, "**ፕረዝደንት**," (*perazedanete*) sounds similar to the English word "*president*."

Moreover, annotators (may) acquire language-specific patterns and rules gradually during annotation; e.g., a capitalized word preceded by "*magaalaa*" is likely to be a city name in Oromo, such as "*magaalaa Adaamaa*" (Adama city). Synchronizing such knowledge among annotators both improves annotation quality and boosts productivity.

The Information Sciences Institute (ISI) developed a "*Chinese Room*" interface[4] to allow a non-native speaker to translate foreign language text into English, based on a small set of parallel sentences that include overlapped words. Inspired by this, RPI and JHU developed two collaborative annotation platforms that exploit linguistic intuitions and resources to allow non-native speakers to perform name tagging efficiently and effectively.

2 Desiderata

We see the following requirements as being most important to allow a non-speaker to annotate a language, independent of interface. None of these requirements is necessary, but the more that are satisfied, the easier it will be for the annotator to produce accurate annotations:

Word recognition. Presentation of text in a familiar alphabet makes it easier to see similarities and differences between text segments, to learn aspects of the target language morphology, and to remember sequences previously seen.

Word pronunciation. Because named entities often are transliterated into another language, access to the sound of the words is particularly important for annotating names. Sounds can be exposed either through a formal expression language such as IPA,[5] or by transliteration into the appropriate letters of the annotator's native language.

Word and sentence meaning. The better the annotator understands the full meaning of the text being annotated, the easier it will be both to identify which named entities are likely to be mentioned in the text and what the boundaries of those mentions are. Meaning can be conveyed in a variety of ways: dictionary lookup to provide fixed meanings for individual words and phrases; description of the position of a word or phrase in a semantic space (e.g., Brown clusters or embedding space) to define words that are not found in a dictionary; and full sentence translation.

Word context. Understanding how a word is used in a given instance can benefit greatly from understanding how that word is used broadly, either across the document being annotated, or across a larger corpus of monolingual text. For example, knowing that a word frequently appears adjacent to a known person name suggests it might be a surname, even if the adjacent word in the current context is not known to be a name.

World knowledge. Knowledge of some of the entities, relations, and events referred to in the text allows the annotator to form a stronger model of what the text as a whole might be saying (e.g., a document about disease outbreak is likely to include organizations like Red Cross), leading to better judgments about components of the text.

History. Annotations previously applied to a use of a word form a strong prior on how a new instance of the word should be tagged. While some of this knowledge is held by the annotator, it is difficult to maintain such knowledge over time. Programmatic support for capturing prior conclusions (linguistic patterns, word translations, possible annotations for a mention along with their frequency) and making them available to the annotator is essential for large collaborative annotation efforts.

[3] https://github.com/andyhu/transliteration

[4] https://www.isi.edu/ ulf/croom/ChineseRoomEditor.html

[5] https://en.wikipedia.org/wiki/IPA

Adjudication. Disagreements among annotators can indicate cases that require closer examination. An adjudication interface is beneficial to enhance precision (see Section 4).

The next section discusses how we embody these requirements in two annotation platforms.

3 Annotation Platforms

We developed two annotation tools to explore the range of ways the desiderata might be fulfilled: ELISA and Dragonfly. After describing these interfaces, Figure 1 shows how they fulfill the desiderata outlined in Table 2.

3.1 ELISA

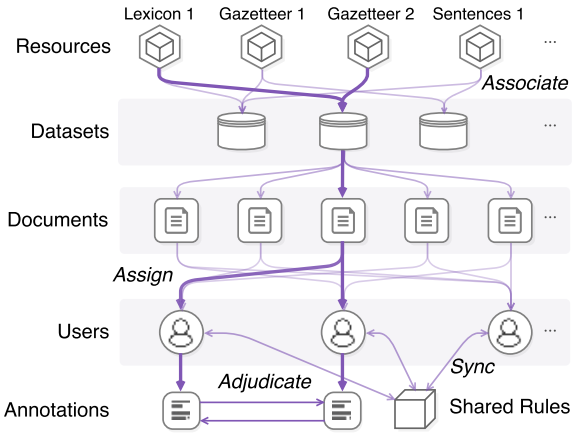

Figure 1: ELISA IE Annotation Architecture.

The ELISA IE annotation platform was developed at Rensselaer Polytechnic Institute.[6] Figure 1 depicts ELISA's overall architecture. Figure 2 demonstrates the main annotation interface, which consists of:

Annotation Panel. For each sentence in a document, we show the text in the original language, its English translation if available, and automatic romanization results generated with a language-universal transliteration library.[7] To label a name mention, the annotator clicks its first and last tokens, then chooses the desired entity type in the annotation panel. If the selected text span has been labeled before, previous annotations are displayed at the bottom of the panel for reference. Annotated mentions are styled differently according to type.

Resource Lookup Panel. This panel is used to browse/search the associated resources. Right clicking a token in the document will show its full definition in lexicons and bilingual example sentences containing that token. A floating pop-up displaying romanization and simple definition appears instantly when hovering over a token.

Rule Editor. Annotators may discover useful hueristics to identify and classify names, such as personal designators and suffixes indicative of locations. They can encode such clues as *rules* in the rule editor. Once created, each rule is rendered as a strikethrough line in the text and is shared among annotators. For example (Figure 1, if an annotator marks "*agency*" as an organization, all annotators will see a triangular sign below each occurrence of this word.

Adjudication Interface. If multiple users process the same document we can consolidate their annotations through an adjudication interface (Figure 3). This interface is similar to the annotation interface, except that competing annotations are displayed as blocks below the text. Clicking a block will accept the associated annotation. The adjudicator can accept annotations from either annotator or accept the agreed cases at once by clicking one of the three interface buttons. Then, the adjudicator need only focus on disputed cases, which are highlighted with a red background.

3.2 Dragonfly

Dragonfly, developed at the Johns Hopkins University Applied Physics Laboratory, takes a more word-centric approach to annotation. Each sentence to be annotated is laid out in a row, each column of which shows a word augmented with a variety of information about that word.

Figure 4 shows a screenshot of a portion of the Dragonfly tool being used to annotate text written in the Kannada language. The top entry in each column is the Kannada word. Next is a Romanization of the word (Hermjakob et al., 2018). The third entry is one or more dictionary translations, if available. The fourth entry is a set of dictionary translations of other words in the word's Brown cluster. (Brown et al., 1992) While these tend to be less accurate than translations of the word, they can give a strong signal that a word falls into a particular category. For example, a Brown cluster containing translations such as "*Paris*," "*Rome*" and "*Vienna*" is likely to refer to a city, even if no translation exists to indicate which city. Finally, if automated labels for the sentence have been generated, e.g., by a trained name tagger, those labels

[6] See examples at http://nlp.cs.rpi.edu/demo/elisa_annotation.html.

[7] https://github.com/andyhu/transliteration

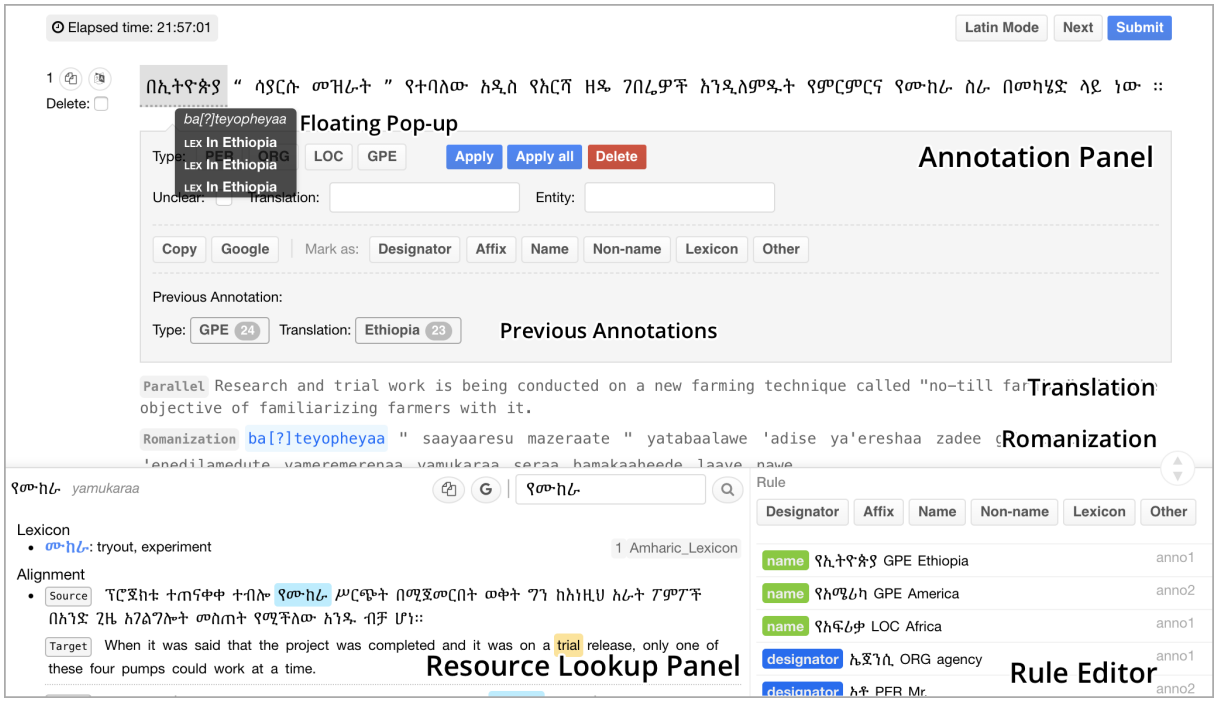

Figure 2: ELISA IE Annotation Interface in use annotating a Tigrinya document.

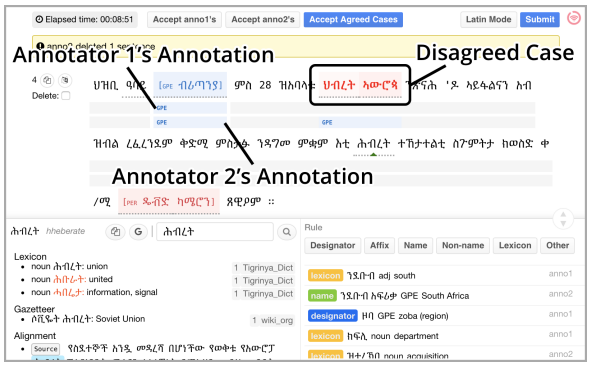

Figure 3: ELISA IE Adjudication Interface in use annotating a Tigrinya document.

are displayed at the bottom of the column.

In addition to word-specific information, Dragonfly can present sentence-level information. In Figure 4, an automatic English translation of the sentence is shown above the words of the sentence (in this example, from Google Translate). Translations might also be available when annotating a parallel document collection. Other sentence-level information that might prove useful in this slot includes a topic model description, or a bilingual embedding of the entire sentence.

Figure 4 shows a short sentence that has been annotated with two name mentions. The first word of the sentence (Romanization "*uttara*") has translations of "*due north*," "*northward*," "*north*," etc. The second word has no direct translations or Brown cluster entries. However, its Romanization, "*koriyaavannu*," begins with a sequence that suggests the word 'Korea' with a morphological ending. Even without the presence of the phrase "*North Korea*" in the MT output, an annotator likely has enough information to draw the conclusion that the GPE "*North Korea*" is mentioned here. The presence of the phrase "*North Korea*" in the machine translation output confirms this choice.

The sentence also contains a word whose Romanization is "*ttramp*." This is a harder call. There is no translation, and the Brown cluster translations do not help. Knowledge of world events, examination of other sentences in the document, the translation of the following word, and the MT output together suggest that this is a mention of "*Donald Trump*;" it can thus be annotated as a person.

4 Experiments

We asked ten non-speakers to annotate names using our annotation platforms on documents in various low-resource languages released by the DARPA LORELEI program and the NIST TAC-KBP2017 EDL Pilot (Ji et al., 2017). The genres of these documents include newswire, discussion forum and tweets. Using non-speaker annotations as "*silver-standard*" training data, we trained

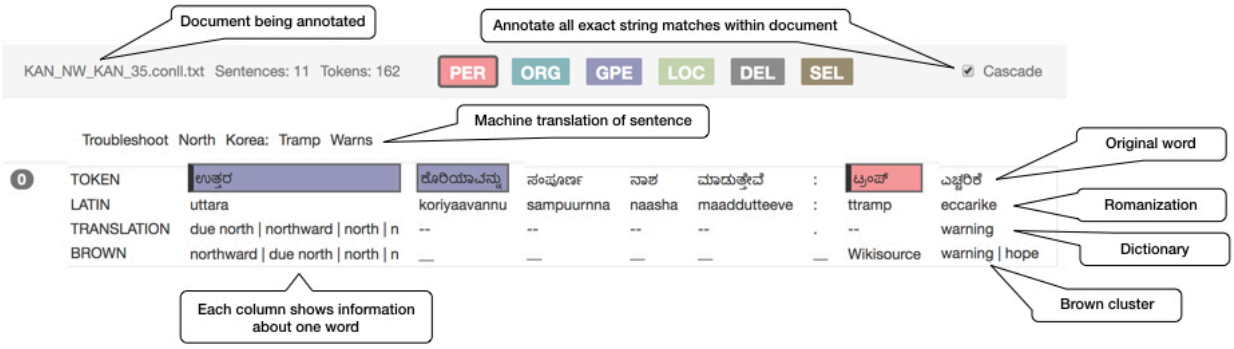

Figure 4: The Dragonfly tool in use annotating a Kannada document.

	ELISA	Dragonfly
Recognition & pronunciation	universal transliteration	uroman
Meanings	Resource Lookup Panel; pop-ups; Annotation Panel	Dictionary; Brown clusters; sentence translation
Word context	Resource Lookup Panel	Concordance
World knowledge	External	External
History	Previous annotations; Rule Editor	Cascade; pop-ups
Adjudication	Adjudication interface	None

Table 1: How Platforms Fulfill Desiderata

name taggers based on a bi-directional long short-term memory (LSTM) network with a Conditional Random Fields (CRFs) layer (Lample et al., 2016). The lexicons loaded into the ELISA IE annotation platform were acquired from Panlex,[8] Geonames[9] and Wiktionary.[10] Dragonfly used bilingual lexicons by (Rolston and Kirchhoff, 2016).

4.1 Overall Performance

The agreement between non-speaker annotations from the ELISA annotation platform and gold standard annotations from LDC native speakers on the same documents is between 72% and 85% for various languages. The ELISA platform enables us to develop cross-lingual entity discovery and linking systems which achieved state-of-the-art performance at both NIST LoreHLT2017[11] and ten languages at TAC-KBP EDL2017 evaluations (Ji et al., 2017).

[8] https://panlex.org/
[9] http://www.geonames.org/
[10] https://www.wiktionary.org/
[11] https://www.nist.gov/itl/iad/mig/lorehlt-evaluations

	Albanian	Kannada	Nepali	Polish	Swahili
#sents	1,652	535	959	1,933	1,714
#tokens	41,785	8,158	16,036	26,924	42,715
#dict entries	96,911	9,931	10,048	644,232	216,323
#names	2,683	900	1,413	1,356	2,769
F_1(%)	**75.9**	**58.4**	**65.0**	**55.7**	**74.2**

Table 2: Data Statistics and Performance on KBP2017 EDL Pilot

Four annotators used two platforms (two each) to annotate 50 VOA news documents for each of the five languages listed in Table 2. Their annotations were then adjudicated through the ELISA adjudication interface. The process took about one week. For each language we used 40 documents for training and 10 documents for test in the TAC-KBP2017 EDL Pilot. In Table 2 we see that the languages with more annotated names (i.e., Albanian and Swahili) achieved higher performance.

4.2 Silver Standard Creation

We compare our method with Wikipedia based silver standard annotations (Pan et al., 2017) on Oromo and Tigrinya, two low-resource languages in the LoreHLT2017 evaluation. Table 3 shows the data statistics. We can see that with the ELISA annotation platform we were able to acquire many more topically-relevant training sentences and thus achieved much higher performance.

Data	Oromo	Tigrinya
ELISA Annotated Training	4,717	6,174
Wikipedia Markup Derived Training	631	152
Gold Standard Unsequestered	2,957	2,201

Table 3: # Sentences in Oromo and Tigrinya Data.

Method	Oromo	Tigrinya
ELISA Annotated	**68.2**	**71.3**
Wikipedia Markup	6.2	2.7

Table 4: Comparison of Silver Standard Creation Methods (F-score %).

4.3 Comparison with Native Speaker Annotations

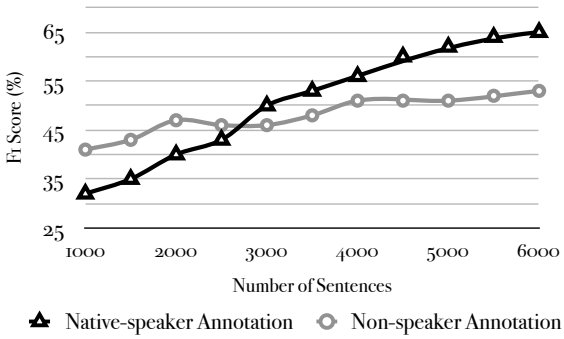

Figure 5: Russian Name Tagging Performance Using Native-speaker and Non-speaker Annotations.

Figure 5 compares the performance of Russian name taggers trained from Gold Standard by LDC native speakers and Silver Standard by non-speakers through our annotation platforms, testing on 1,952 sentences with ground truth annotated by LDC native speakers. Our annotation platforms got off to a good start and offered higher performance than annotations from native speakers, because non-speakers quickly capture common names, which can be synthesized as effective features and patterns for our name tagger. However, after all low-hanging fruit was picked, it became difficult for non-speakers to discover many uncommon names due to the limited coverage of lexicon and romanization; thus the performance of the name tagger converged quickly and hits an upper-bound. For example, the most frequently missed names by non-speakers include organization abbreviations and uncommon person names.

4.4 Impact of Adjudication

Table 5 shows that the adjudication process significantly improved precision because annotators were able to fix annotation errors after extensive discussions on disputed cases and also gradually learned annotation rules and linguistic patterns. Most missing errors remained unfixed during the adjudication so the recall was not improved.

Language	Adjudication	P (%)	R (%)	F (%)
Oromo	Before	68.6	61.3	64.7
	After	**76.2**	**61.8**	**68.2**
Tigrinya	Before	67.3	67.1	67.2
	After	**76.4**	66.8	**71.3**

Table 5: Impact of Annotation Adjudication

References

J. An, S. Lee, and G. G. Lee. 2003. Automatic acquisition of named entity tagged corpus from world wide web. In *ACL*.

P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *COLING* .

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *NAACL HLT 2010 Workshop*.

U. Hermjakob, J. May, and K. Knight. 2018. Out-of-the-box universal romanization tool uroman. In *PRoc. ACL2018 Demo Track*.

H. Ji, X. Pan, B. Zhang, J. Nothman, J. Mayfield, P. McNamee, and C. Costello. 2017. Overview of TAC-KBP2017 13 languages entity discovery and linking. In *TAC*.

G. Lample, M. Ballesteros, K. Kawakami, S. Subramanian, and C. Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT*.

J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence* 194:151–175.

X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *ACL*.

Leanne Rolston and Katrin Kirchhoff. 2016. Collection of bilingual data for lexicon transfer learning. In *UWEE Technical Report*.

C. Tsai, S. Mayhew, and D. Roth. 2016. Cross-lingual named entity recognition via wikification. In *CoNLL*.

M. Wang and C. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics* 2:55–66.

B. Zhang, X. Pan, Y. Lin, T. Zhang, K. Blissett, S. Kazemi, S. Whitehead, L. Huang, and H. Ji. 2017. RPI BLENDER TAC-KBP2017 13 languages EDL system. In *TAC*.