Error Analysis of Uyghur Name Tagging: Language-specific Techniques and Remaining Challenges

Abudukelimu Halidanmu[1], Abulizi Abudoukelimu[1], Boliang Zhang[2], Xiaoman Pan[2], Di Lu[2], Heng Ji[2], Yang Liu[1]

[1] Tsinghua University, [2] Rensselaer Polytechnic Institute

[1] {abdklmhldm}@gmail.com, [1] {keram1106}@163.com, [1] {liuyang2011}@tsinghua.edu.cn,

[2] {zhangb8,panx2,lud2,jih}@rpi.edu

Abstract

Regardless of numerous efforts at name tagging for Uyghur, there is limited understanding on the performance ceiling. In this paper, we take a close look at the successful cases and perform careful analysis on the remaining errors of a state-of-the-art Uyghur name tagger, systematically categorize challenges, and propose possible solutions. We conclude that simply adopting a machine learning model which is proven successful for high-resource languages along with language-independent superficial features is unlikely to be effective for Uyghur, or low-resource languages in general. Further advancement requires exploiting rich language-specific knowledge and non-traditional linguistic resources, and novel methods to encode them into machine learning frameworks.

Keywords: Low-resource Languages, Name Tagging, Error Analysis

1. Introduction

Uyghur is a language spoken by 8.2 million people, primarily by the Uyghur people in the Xinjiang Uyghur Autonomous Region of Western China. In terms of the number of native speakers, it's ranked at the 94th among all the languages in the world [1], but it has extremely low linguistic resources. There are very few Natural Language Processing (NLP) tools, standard annotated corpora, or language universal resources (e.g., World Atlas of Linguistic Structure (WALS) database (Haspelmath et al., 2005; Dryer and Haspelmath, 2013)) available. Even for naturally existing noisy annotations such as Wikipedia markups, Uyghur is ranked very low (the 195th [2]). There are only 2,566 Uyghur pages in Wikipedia, much fewer than its related languages such as Turkish (277,547 pages) and Uzbek (128,664 pages). Most Uyghur Wikipedia pages contain much less content than their counterparts in Turkish, Uzbek and English. The cross-lingual links are not carefully validated and thus contain many errors.

It's certainly important to develop automatic NLP tools for Uyghur, so as to distill information from textual documents written in Uyghur, as well as preserve their unique culture, music, art and the long history of which the Uyghurs are deeply prideful of. Unfortunately, the striking fact is that very little Uyghur NLP work has been published to catch the attention of the wider international NLP research community.

Using Uyghur name tagging as a case study, some previous studies (Li et al., 2011; Arkin et al., 2013b; Rozi et al., 2013; Turhun et al., 2012; Li, 2014; Arkin et al., 2013c; Arkin et al., 2013a; Maihefureti et al., 2014; Zhang, 2014; Zhang et al., 2015; Yu et al., 2015; Nizamidin et al., 2016) have adopted popular machine learning methods which were effective for other high-resource languages. The features suggested by these previous papers include: numbers, shape, stem, suffix, the number of suffixes, first syllable, last syllable, the number of syllables, Part-of-speech tags, the closest verb, word length, position in the sentence and special rules to identify Chinese person names. Further advances in this field require us to look into language-specific problems and recommended solutions to those challenges.

In this paper, we will look at the remaining errors of a high-performing Uyghur name tagger, and decompose the remaining errors into detailed categories in order to understand how varied components may contribute to improvement. We believe such comprehensive, quantitative and qualitative error analysis may help draw a roadmap for future research and resource development on this important and yet challenging task.

2. Approach Overview

We use a deep neural networks based Uyghur name tagger as our target system for analysis because of two reasons: (1) it achieves top performance at NIST LoreHLT2016 Evaluation [3] so it represents state-of-the-art; (2) unlike most previous work, this system has already exploited extensive language-specific features. We briefly describe the system as follows.

2..1 Learning Model

This system considers 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 (Person (PER), Organization (ORG), Geo-political Entity (GPE), and Location (LOC)). Following a framework similar to (Lample et al., 2016). The architecture consists of Bi-directional Long Short-Term Memory and Conditional Random Fields (CRFs) network. After processing through the Bi-LSTM networks, each token in a sentence sequence obtains a feature embedding that captures left and

[1] https://en.m.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] https://www.nist.gov/itl/iad/mig/lorehlt16-evaluations

right context information, which is then fed into the CRF networks.

2..2 Pre-processing

The system starts with segmenting a document into tokens based on 50 punctuations pulled from Uyghur grammar books ((translated by Anne Lee), 2003; Zakir, 2007; Engesæth et al., 2009). Since Uyghur is a morphologically rich language, a set of name related suffixes is also extracted from grammar books, Wikitionary [4] and WALS [5], for stemming and feature encoding.

2..3 Features

Typical **implicit** linguistic features including character embeddings and word embeddings are learned from a large monolingual unlabeled corpus from LoreHLT2016 and then fed into the Bi-LSTM networks. Moreover, the following Uyghur-specific **explicit** context-dependent linguistic features are directly fed into the final CRFs model.

- The first and the last syllables of each token, based on the intuition that Uyghur names often include suffixes, and the first syllables of person names often follow some specific patterns.
- 319 common syllable patterns from person names. The most frequent patterns include (Latin: gue), (Latin: sha), مىر (Latin: mir), غا (Latin: gha), (Latin: uel), (Latin: ash), گۇل (Latin: guel), (Latin: buew), ئابدۇ (Latin: abdu) and مۇھەممە (Latin: muhemme)
- Suffixes are categorized into three types of features: (1) indicating animacy so that the word is likely to be part of a person or an organization name, including نىڭ (Latin: ning), نى (Latin: ni), لۇق (Latin: luq), and لىك (Latin: lik). (2) locative suffixes indicating GPE/LOC names, including كە (Latin: ke), گە (Latin: ge), قا (Latin: qa), غا (Latin: gha), تە (Latin: te), دە (Latin: de), تا (Latin: ta), دا (Latin: da), تىن (Latin: tin), دىن (Latin: din), تىكى (Latin: tiki), دىكى (Latin: diki), كىچە (Latin: kiche), گىچە (Latin: giche), قىچە (Latin: qiche), and غىچە (Latin: ghiche). (3) suffixes indicating a word is unlikely to be a name or part of a name, including لار (Latin: lar) and لەر (Latin: ler).
- Two words before and two words after the current token.
- Conjunction feature of stem and suffix.
- Name designators: English name designators are translated into Uyghur through a bi-lingual lexicon from LoreHLT2016.
- 446 Chinese last names are translated into Uyghur.

2..4 Post-processing

In the low-resource setting, the available resources are not sufficient to generalize some Uighur-specific linguistic phenomena as in high-resource language setting. We designed the following heuristic rules as post-processing to fix some obvious errors in informal genres like discussion forum and tweets.

- Remove a name if it includes digital numbers but it's not a poster or Twitter ID.
- If a name includes a URL link, remove the URL.
- Label places that don't have governing organizations as Location(LOC), including all continents, ئوتتۇرا شەرق (Latin: ottura sheriq, English: Middle East), etc.
- Label places with location modifiers as LOC, e.g., جەنۇبىي ئامرىكا (Latin: jenubiy amirika, English: North America).
- Label countries of countries as GPE, e.g., ياۋروپا ئىتتىپاقى (Latin: yawrupa ittipaqi, English: the European Union).
- Remove generic name mentions of people of the certain ethnicity (e.g., 'Uighur People', 'Americans', 'Arabs') by checking the combinations of country names and suffixes indicating 'people'.
- There are many very long nested organizations whose boundaries are difficult to determine. The basic principle is to tag every different, distinct entity by checking if it should be created as a unique entry when we construct a knowledge base. For example, بىرلەشكەن دۆلەتلەر تەشكىلاتى مائارىپ، پەن-مەدەنىيەت تەشكىلاتىنىڭ جۇڭگودا تۇرۇشلۇق ۋەكىللەر ئۆمىكى ئۇچۇر تارقىتىش باشقارمىسىنىڭ ("*Media Communication office from the United Nations Educational, Scientific, and Cultural Organization in China*") should be tagged as one single organization mention.
- Boundary extension: if a name doesn't include any suffix and its right contextual word is a name designator, then extend the name boundary to include the designator.
- Cross-genre propagation: when the types of the same name mention are conflicting between formal genres and informal genres, propagate the types from formal genres to informal genres.
- Poster names: Extract all poster names from the original thread structures, and identify all mentions in the posts, posters, and Twitter user names. Apply English entity linking (Pan et al., 2015) to each string after '@' or '#', and if it's linkable and its type can be inferred based on KB properties, then assign the type; otherwise tag it as PER.

2..5 Performance

For the experiment in this paper, we used the unsequestered Uyghur documents from the NIST LoReHLT16 evaluation. We used 99 documents for training and 30 documents for test and achieved 65.23% F-score. This performance is encouraging given the limited resources. The above explicit linguistic features achieved 2.4% F-score improvement. However, the overall performance is still much lower than other languages such as English, Spanish and Dutch (Lample et al., 2016), and also much lower than Uzbek (close to 80% F-score trained from a similar size of data) which is a similar language as Uyghur but has much more linguistic resources. In next section we will take a close look at the remaining errors.

3. Error Analysis

One major challenge to develop NLP techniques for low-resource languages is that system developers usually have little knowledge about the languages so it's very difficult to

[4] https://en.wiktionary.org/wiki/Wiktionary:Main_Page
[5] http://wals.info/

perform effective error analysis in order to do hill-climbing. Thus in this paper, we ask two Uyghur native speakers to focus on detailed error analysis. In this section, we aim to explain why Uyghur name tagging is so challenging, and discuss various methods we attempted to address these challenges, and potential language-specific solutions.

3..1 Error distribution

In Figure 1 we present the distribution of errors which need different techniques, according to their difficulty levels. The percentage numbers are approximate because some errors may rely on the combination of multiple types of features.

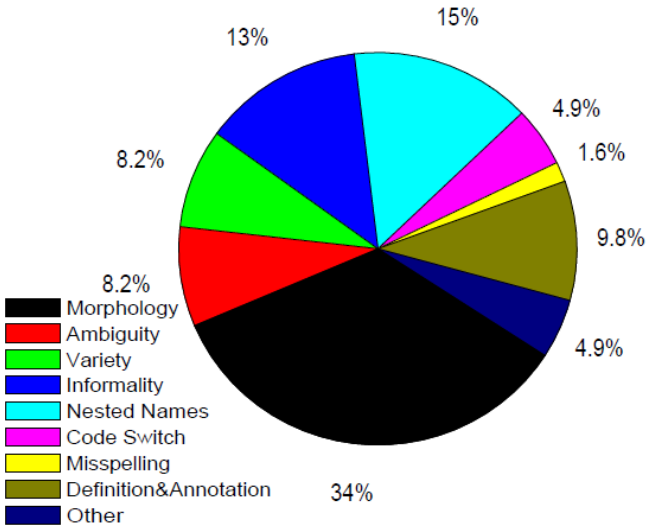

Figure 1: Error Type Distribution

3..2 Rich Morphology

From Figure 1 we can clearly see that morphology analysis dominates the causes for errors. Uyghur is a very 'sticky' language that contains very rich morphology. 90% words have arbitrary combinations of multiple suffixes denoting person, number, case, mood, etc. Sometimes a compound Uyghur word may even indicate an entire sentence, as shown in Table 1. Due to a large number of suffix combinations, among all of the unique tokens in our corpus, only 68.1% of them exist in the LDC provided lexicon, while 31.9% are Out-of-Vocabulary (OOV) words.

More than 90% GPE names include suffixes. State-of-the-art language universal morphology analyzers such as Morfessor (Creutz, 2003) don't perform well on identifying Uyghur suffixes for name tagging purpose. The suffixes we mined from grammar books, Wikitionary and WALS can perform some reasonable amount of stemming. However, it still remains a big challenge for machine learning models to generalize these rich suffix combinations.

3..3 Ambiguity

Names and their contexts are highly ambiguous in Uyghur. We further categorize the ambiguous cases as follows.

Ambiguity between name and non-name. A Uyghur person or GPE name usually has a positive meaning, and thus it can also be a common word (usually noun or adjective) in different contexts. Table 2 shows some examples.

Table 1: Uyghur Suffix Derivation Examples

Uyghur	English
زامان zaman	era
زامانۋى zamanivi	modern
زامانۋىلاش zamanivilax	modernization
زامانۋىلاشتۇر zamanivilaxtur	modernize
زامانۋىلاشتۇرۇل zamanivilaxturur	be modernized
زامانۋىلاشتۇرۇلما zamanivilaxtururma	can not be modernized
زامانۋىلاشتۇرۇلمايمىز zamanivilaxtururmaimiz	We can not be modernized

Table 2: Ambiguous Uyghur Name Examples

Type	Name	Meaning as Non-Name
PER	ئارسلان	lion
	ئالىم	scientist
	روزا	fasting
	قەھرىمان	hero
	بەختيار	happy
	ھېيت قۇربان	Corban Festival
	ئىسلام	Islam
	گۈزەل	beatiful
	يولۋاس	tiger
	تۇردى	stand
	رەجەپ	July
	ئاينۇر	moonlight
	جىنەستە	cherry
	دىلدار	sweetheart
	رەيھان	violet
	نىگار	lover
	ئايتۇرۇك	Moon of the Turkish
GPE	ئادىل	fair
	ئارال	island (transliterated as 'Alar')
	دۆلەت	country (transliterated as person name 'Dolet')

Ambiguity across name types. There also exists a lot of ambiguity across name types. Table 3 presents some examples of name mentions indicating various types in different contexts.

To solve these problems, we will need to develop and exploit more language-specific resources such as title lists and tools such as dependency parser to capture wider contexts. This ambiguity problem also makes it difficult to transfer indicative contextual words in English name tagger to Uyghur. On average any English action verb or title has more than 10 possible translations in Uyghur. For example, the word "*watch*" has 41 possible translations in Uyghur.

Table 3: Ambiguity across name types Ambiguous

Name	Translation	Type1	Type2
سايرام	Sayram	LOC	PER
ئالىم	scientist	PER	Nominal
نەبىجان دۆلەت	Nebijan country	GPE	PER
ئارمان	ARMAN	PER	ORG
ئابىدە	ABIDE	PER	ORG
ئىخلاس	IHLAS	PER	ORG
جۇڭغار	Dzungaria	LOC	ORG

Table 4: Informal Names

Uyghur Name	Literal Translation	Referent Entity
تىرامپ ھۆكۈمىتى	Trump government	USA
شام دۆلىتى	candle country	Islamic State
قاراقاش	thick eyebrows	Karakax County
تىبەت	Tibetan	Tibet
شىنزۇ ئابى ھۆكۈمىتى	Abe Shinzo government	Japan government
ئاستانە	Capital	Astane County

3..4 Variety

Name variety. Names with different origins (Uyghur, Han Chinese, foreign names) have different characteristics. For example, a place name in Xinjiang can be transliterated either by Chinese pinyin or by its original Uyghur pronunciation. For example, both "*Hetian*" and "*Hotan*" refer to the same city that appear frequently in English news articles. There is no established standard yet for transliterating Uyghur names, which makes it difficult to project name gazetteers in high-resource languages such as English or Chinese onto Uyghur documents for name identification. After romanizing Uyghur, many foreign names look very similar to their English forms. For example, "*donald trumpni*" refers to "*Donald Trump*", "*nato ken*" refers to "*Naoto Kan*", and "*amerika*" refers to "*America*". Therefore we tried to use a Soundex based matching method to perform cross-lingual entity linking on each romanized Uyghur ngram to English Knowledge Base and English gazetteers in order to determine whether it's a name. However, this simple approach produced many spurious errors. In the future, it might be more effective to add it as an implicit feature in the model.

Unlike English, there is no capitalization for names in Uyghur. For Han Chinese person names, there is a fixed list of last names which are usually one single character, and each first name is usually a limited 1-2 character. However, neither of these two characteristics exists in Uyghur person names. Similar to Turkish, a Uyghur person's last name is his/her father's first name. Moreover, each first name or last name is usually a common word that carries some positive meaning, which yields an almost infinite number of combinations. Therefore it's more challenging to determine Uyghur person name boundaries than English, Chinese and Turkish.

Context Variety. We also attempted to project English word embeddings to Uyghur using a bilingual lexicon. However, the available lexicon has too low coverage to provide any gains. The same approach provided significant gains (up to 5%) for both Turkish and Uzbek name tagging.

3..5 Informal Names

Due to historical and cultural reasons, a substantial amount of informal Uyghur names is being created, especially at social media platforms. Table 4 presents some examples. Our model also missed some informal poster names and twitter users which don't appear in indicate contexts. Many of these names are common words such as سەپەر ("*travel*"). In addition, many Uighur people like to create pen names for themselves as their middle names, which are also difficult to identify.

3..6 Long Nested Organizations

Organization names in Uyghur texts are often very long containing nested names. Some challenging examples are presented in Table 5. It's difficult to determine their boundaries, especially when they are not linkable to external knowledge bases, or contain names which are also common words (e.g., the nested person name "*Arman*" means "*dream*"; and "*Abida*" means "*milestone*"). Addressing this challenge would require us to develop more advanced name internal structure parsing techniques.

3..7 Code Switch

Another unique challenge of Uyghur name tagging is the frequent code-switch phenomena in Uyghur texts. A large variety of names are borrowed from other languages, including Mandarin which is taught in most Uyghur schools (e.g., جۇڭگو ("*China*"), شىخەنزە ("*Shihezi*"), تاۋباۋ ("*Taobao*")), Arabic which is due to religious reasons (e.g., مۇھەممەت ("*Mohammed*")), English (e.g., ئامېرىكا ("*United States*"), تىرامپ ("*Trump*")) and Russian which are due to commercial trades. Extracting these names correctly requires us to identify their origins and capture the detailed characteristics on how they were transliterated.

3..8 Misspellings

Many names and contextual words in Uyghur texts in both formal and informal genres are misspelled. For example, the common person name دېڭ شياۋپىڭ (*Deng Xiaoping*) is often misspelled as دېڭ شياۋپېڭ even including its Wikipedia title; the Wikipedia title of جۇڭگو ("*China*") is also misspelled as جوڭگو; ئۈرۈمچى (*Urumqi*) is often misspelled as ئۇرۇمچى. Up to date there are no effective Uyghur spelling correction techniques available yet.

3..9 Name Definition and Annotation Challenges

In the past two decades, many efforts have been made at defining the name tagging task, including Message Understanding Conference (MUC) (Grishman and Sundheim, 1996), Automatic Content Extraction (ACE) [6], and Entity, Relation and Event (ERE) (Song et al., 2015). However, there are many open issues which may cause confusions for both human annotators and systems. In particular, for low-resource languages like Uyghur, it's also challenging to train native speakers to follow a long annotation guideline

[6] http://www.itl.nist.gov/iad/mig/tests/ace/

Table 5: Long Nested Organizations

Nested Organization	English Translation
[ORG مۇسۇلمانچە يىمەكلىك سانائىتى گۇرۇھى چەكلىك شىركىتى [PER ئارمان] [GPE شىنجاڭ]]	[ORG [GPE Xinjiang] [PER Arman] Halal Food Group Co., Ltd.]]
[ORG [ئابدەPER] بىئو پەن - تېخنىكا تەرەققىيات چەكلىك شىركىتى [GPE شىنجاڭ]	[ORG [GPE Xinjiang] [PER Abida] Biotechnology Development Co., Ltd.]
[ORG [ORG [GPE ئېكولوگىيەسى ۋە جۇغراپىيە تەتقىقات مەركىزىنىڭ [GPE شىنجاڭ]] [ORG جۇڭگو پەنلەر ئاكادېمىيەسى]]	[ORG [ORG [GPE Xinjiang] Institute of Ecology and Geography] [ORG Chinese Academy of Sciences]]
[ORG [ORG مۇخبىرلار پونكىتى [GPE شىنجاڭ]] [ORG شىنخۇا ئاگېنتلىقى]]	[ORG [ORG [GPE Xinjiang] Editorial Office] of [ORG Xinhua News Agency]]

(usually more than 20 pages) which may still leave many language-specific issues underspecified or unresolved. Two annotators at Linguistic Data Consortium (LDC) performed independent annotations on a subset of the LoreHLT2016 Uyghur name tagging data. Compared to the ground truth their F-scores were only 60.5% and 78.8% respectively. In the following, we will discuss some remaining gray areas which may still lead to confusions and different interpretations, and thus often it's difficult to draw a line. These problems are often amplified due to the communication barrier among the guideline developers, system developers and annotators.

- Adjective form and roles: Names in adjectival form, or modifiers, such as "*[GPE American] army*", are taggable. But this definition causes confusions because the modifiers do not always play geo-political entity roles (e.g., "*Chinese*" in "*[GPE Chinese] food*"). On the other hand, when news organization names refer to publications instead of organization roles, they should not be tagged. For example, in "*Bob enjoys reading the New York Times*", "*New York Times*" should not be tagged as an organization. Similarly, when a facility (e.g., "*White House*") plays an ORG role (e.g., make a statement), it should be tagged as ORG. Accurately determining these semantic roles requires further deep understanding of implicit contexts.
- Designator: it's often debatable whether GPE/LOC/ORG designators like "*city*" and "*company*" should be included in name mentions.
- Specific entity: Most guidelines indicate that names of deceased people, fictional characters, religious entities should all be tagged. In contrast generic persons are not taggable, such as "*Americans*", "*Christians*", "*Arabs*", "*Democrats*" and "*Republicans*".
- Group entity: When a GPE name is used to refer to the people of a GPE, it should not be tagged as a PER or GPE name. For example, in "*The Swiss have joined us on the bus tour*", "*Swiss*" should not be tagged. In contrast, a group of countries such as "*the European Union*" should be tagged as GPE.
- Entity subpart: a subpart of GPE that doesn't have a government (e.g., "*South America*", "*North America*", "*Middle East*", "*South Asia*") should be tagged as a LOC.
- Nominal mentions: when a nominal mention refers to a specific entity with rich context, both human annotators and systems tend to mistakenly label it as a name. For example, in the following sentence "*According to the report from China News Web, the telegraph from Xinhua Network's Bureau in Xinjiang stated that after the earthquake in Kiriye, four teams, consisting of the members from the local army, the civil official, the health department, the police, the fire department, the power supply and so on, have left for the affected Atchan township.* ", it's difficult to decide whether "*the health department*" and "*the fire department*" are names or nominals.

4. Conclusions and Future Work

We conducted a thorough study on both quantitative and qualitative analysis on a wide variety of errors from a state-of-the-art Uyghur name tagger. We also discussed possible solutions for the remaining challenges. Recently there is a trend in the community to push the rapid development of language universal techniques for name tagging (Zhang et al., 2016; Littell et al., 2016; Tsai et al., 2016; Pan et al., 2017). These methods have achieved some success at setting up baseline name taggers with reasonable performance. However, based on the Uyghur case study in this paper we can clearly see that most of the remaining challenges are specific to the target language, and thus we will need to embrace language-specific resources and knowledge in order to break the performance ceiling. We hope that the detailed analysis we did in this paper can shed a light on future efforts to focus on Uyghur resource development instead of simply borrowing language-independent features and machine learning methods which were used by other languages.

References

Arkin, M., Hamdulla, A., and Tursun, D. (2013a). Recognition of uyghur place names based on rules. *Communications Technology*, 7.

Arkin, M., Mahmut, A., and Hamdulla, A. (2013b). Person name recognition for uyghur using conditional random fields. *IJCSI International Journal of Computer Science Issues*, 10.

Arkin, M., Mahmut, A., and Hamdulla, A. (2013c). Person name recognition for uyghur using conditional random fields. *International Journal of Computer Science Issues*, 10.

Creutz, M. (2003). Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL2003*.

Matthew S. Dryer et al., editors. (2013). *WALS Online*.

Engesæth, T., Yakup, M., and Dwyer, A. (2009). *Greetings from the Teklimakan: a Handbook of Modern Uyghur*.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Proceedings of COLING1996*.

Martin Haspelmath, et al., editors. (2005). *World Atlas of Language Structures*.

Lample, G., Ballesteros, M., Kawakami, K., Subramanian, S., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proc. the 2016 Conference of the North American Chapter of the Association for Computational Linguistics —Human Language Technologies (NAACL-HLT 2016)*.

Li, J., Liu, K., Aili, M., Lv, Y., Liu, Q., and Yibulayin, T. (2011). Recognition and translation for chinese names in uyghur language. *Journal OF Chinese Information Processing*, 25.

Li, J. (2014). Research on uyghur named entity recognition and translation.

Littell, P., Goyal, K., Mortensen, R. D., Little, A., Dyer, C., and Levin, L. (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*.

Maihefureti, Rouzi, M., Aili, M., and Yibulayin, T. (2014). Uyghur organization name recognition based on syntactic and semantic knowledge. *Computer Engineering and Design*, 35.

Nizamidin, T., Tuerxun, P., Hamdulla, A., and Arkin, M. (2016). A survey of uyghur person name recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9.

Pan, X., Cassidy, T., Hermjakob, U., Ji, H., and Knight, K. (2015). Unsupervised entity linking with abstract meaning representation. In *Proc. the 2015 Conference of the North American Chapter of the Association for Computational Linguistics —Human Language Technologies (NAACL-HLT 2015)*.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)*.

Rozi, A., ZONG, C., Mamateli, G., Mahmut, R., and Hamdulla, A. (2013). Approach to recognizing uyghur names based on conditional random fields. *Journal of Tsinghua University Science and Technology*, 53.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015*.

(translated by Anne Lee), H. T. (2003). *Modern Uyghur Grammar (Morphology)*.

Tsai, C.-T., Mayhew, S., and Roth, D. (2016). Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Turhun, N., Yu, H., and Li, Y. (2012). Research on crf based uyghur named entity recognition. *Consumer Electronics Magazine*, 7.

Yu, G., Xu, Y., Ma, M., Yu, Z., and Arkin, M. (2015). Research on rule based uyghur named entity recognition. *China Science and Technology Panorama Magazine*, 15.

Zakir, H. A. (2007). *Introduction to Modern Uighur*.

Zhang, L., Yang, Y., Mi, C., and Li, X. (2015). Recognition and translation of uyghur named entities in numerals class. *Computer Applications and Software*, 32.

Zhang, B., Pan, X., Wang, T., Vaswani, A., Ji, H., Knight, K., and Marcu, D. (2016). Name tagging for low-resource incident languages based on expectation-driven learning. In *Proc. the 2016 Conference of the North American Chapter of the Association for Computational Linguistics —Human Language Technologies (NAACL-HLT 2016)*.

Zhang, L. (2014). Recognition and translation of uyghur name entities for chinese-uyghur machine translation.