International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.3 (2016), pp.273-280
http://dx.doi.org/10.14257/ijsip.2016.9.3.24

A Survey of Uyghur Person Name Recognition

Tashpolat Nizamidin[1], Palidan Tuerxun[2], Askar Hamdulla[2*] and Muhtar Arkin[3]

[1]*Institute of Information Science and Engineering, Xinjiang University, China*
[2]*School of Software, Xinjiang University, Urumqi, China 830046,*
[3]*College of Information Science and Engineering, Urumqi Vocational University, Urumqi, Xinjiang, China*
[1]*tashifulati@qq.com,* [2]*askarhamdulla@sina.com,* [3]*muhtararkin@yahoo.com*

Abstract

Uyghur is one of the most populous and civilized groups with Turkic ethnicity and mainly located Xinjiang Uyghur Autonomous Region of China. Uyghur language belongs to the Karluk branch of the Turkic language family in Altaic language system, and holds agglutinative characteristics in morphological structure. Named Entity Recognition (NER) is an Information Extraction task that has become an essential part of Natural Language Processing (NLP) tasks, such as Machine Translation and Information Retrieval. In this paper, as a subtask of NER, the importance of Uyghur Named Entity Recognition (UPNR) task is demonstrated, the main characteristics of the Uyghur language are highlighted, and the aspects of standardization in annotating named entities are illustrated. Moreover, the approaches used in Uyghur NPNR field are explained and the features of common tools used in Uyghur NPNR are described. A brief review of the state of the art of Uyghur NPNR research is discussed, too. Finally, we present our conclusions. Throughout the presentation, illustrative examples are used for clarification.

Keywords: *Uyghur, NER, Person Name Recognition, Machine learning*

1. Introduction

The term "Named Entity", now widely used in Natural Language Processing, was coined for the Sixth Message Understanding Conference (MUC-6) (R. Grishman & Sundheim 1996) [1]. At that time, MUC was focusing on Information Extraction (IE) tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units like names, including person, organization, location names and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition" .

Person, organization and location names have different characteristics in Uyghur language. Therefore, various entities were studied respectively. This paper mainly introduces the study of Person Name Recognition in Uyghur. We present here a survey of five years of research in UPNR, from 2011 to 2015. While early systems were making use of handcrafted rule-based algorithms, modern systems most often resort to machine learning techniques. We survey these techniques as well as other critical aspects of UPNR such as features and evaluation methods. It was indeed concluded in a recent conference that the choice of features is at least as important as the choice of technique for obtaining a good NER system.

The following section of this survey presents some observations on published work from the point of yearly activities. It was collected from the review of several Uyghur

* Corresponding Author

ISSN: 2005-4254 IJSIP

Person Name Recognition papers sampled from the major conferences and journals in this field. Section 3 covers the algorithmic techniques that were proposed for addressing the NER task. Instead of elaborating on techniques themselves, the fourth section lists and classifies the proposed features, *i.e.,* characteristics of Uyghur words for UPNR. Finally, in the last section, we present our conclusions.

2. Observations: 2011-2015

The computational research aiming at automatically identifying Person Names in texts forms a vast and heterogeneous pool of strategies, methods and representations. One of the first research papers in the field was presented by Li Jiazheng (2011) at The Journal of Chinese Information. Li's paper describes a method for recognizing and translating Chinese names in Uyghur. It relies on both Uyghur and Chinese language models, in addition to using the traditional rule-based approach.

The first statistical approach for Uyghur Person Name Recognition was presented by Askar Rozi (2013) at The Journal of Tsinghua University (Science and Technology) [2]. In his paper, conditional random fields (CRFs) was applied for recognizing Uyghur person names The agglutinative characteristics of Uyghur language were used to determine the word, part-of-speech, word stem, suffix, first and last syllables, and the nearest verb as features. A greedy algorithm was used to select the best feature templates for the recognition model.

Jarulla Muhammad, *et al.,* (2014) presented a hybrid approach for automatic identifying of Uyghur person names at The Journal of Xinjiang University (Natural Science Edition) [3]. This method realized identifying of candidate person names and eliminated ambiguity using statistical boundary model through analyzing the characteristics of Uygur persons names, extracted feature sets and summarization of corresponding recognition rules

Hussein Yusuf, *et al.,* (2015) presented a rule-based approach at the National Conference on Man-Machine Speech Communication which uses Dice Coefficient and Levenshtein Distance. A letter-based fuzzy matching method was used for Uyghur person names, while the syllable-character conversion method which is inspired by the idea of machine translation was used for Chinese person names.

3. Learning Methods

3.1. Supervised Learning

The current dominant technique for addressing the NER problem is supervised learning. SL (Supervised Learning) techniques include Hidden Markov Models (HMM) (D. Bikel, *et al.,* 1997) [4], Decision Trees (S. Sekine 1998) [5], Maximum Entropy Models (ME) (A.Borthwick 1998) [6], Support Vector Machines (SVM) (M. Asahara & Matsumoto 2003) [7], and Conditional Random Fields (CRF) (A. McCallum & Li 2003) [8]. These are all variants of the SL approach that 5 typically consist of a system that reads a large annotated corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features.

A baseline SL method that is often proposed consists of tagging words of a test corpus when they are annotated as entities in the training corpus. The performance of the baseline system depends on the vocabulary transfer, which is the proportion of words, without repetitions, appearing in both training and testing corpus. D. Palmer and Day (1997) calculated the vocabulary transfer on the MUC-6 training data. They report a transfer of 21%, with as much as 42% of location names being repeated but only 17% of organizations and 13% of person names. Vocabulary transfer is a good indicator of the recall rate (number of entities identified over the total number of entities) of the baseline system, but is a pessimistic measure since some entities are frequently repeated in

documents. A. Mikheev, *et al.,* (1999) precisely calculated the recall of the baseline system on the MUC-7 corpus. They report a recall of 76% for locations, 49% for organizations and 26% for persons with precision ranging from 70% to 90%. White Law and Patrick (2003) report consistent results on MUC-7 for the aggregated enamex class. For the three enamex types together, the precision of recognition is 76% and the recall rate is 48%.

3.2. Semi Supervised Learning

The term "semi-supervised" (or "weakly supervised") is relatively recent. The main technique for SSL (Semi supervised learning) is called "bootstrapping" and involves a small degree of supervision, such as a set of seeds, for starting the learning process. The learning process is then reapplied to the newly found examples, so as to discover new relevant contexts. By repeating this process, large number of contexts will eventually be gathered. Recent experiments in semi-supervised NER (Nadeau, *et al.,* 2006) report performances that rival baseline supervised approachesm [9]. Here are some examples of SSL approaches. S. Brin (1998) uses lexical features implemented by regular expressions in order to generate lists of book titles paired with book authors [10]. It starts with seed examples such as {Isaac Asimov, The Robots of Dawn}. The main idea of his algorithm, however, is that many web sites conform to a reasonably uniform format across the site. When a given web site is found to contain seed examples, new pairs can often be identified using simple constraints such as the presence of identical text before, between or after the elements of an interesting pair. For example, the passage "The Robots of Dawn, by Isaac Asimov (Paperback)" would allow finding, on the same web site, "The Ants, by Bernard Werber (Paperback)".

M. Collins and Singer (1999) [11] parse a complete corpus in search of candidate NE patterns. A pattern is, for instance, a proper name (as identified by a part-of-speech tagger) followed by a noun phrase in apposition (*e.g.,* Maury Cooper, a vice president at S&P). Patterns are kept in pairs {spelling, context} where spelling refers to the proper name and context refers to the noun phrase in its context. Starting with an initial seed of spelling rules, the candidates are examined. Candidate that satisfy a spelling rule are classified accordingly and their contexts are accumulated. The most frequent contexts found are turned into a set of contextual rules. Following the steps above, contextual rules can be used to find further spelling rules, and so on. M. Collins and Singer and R. Yangarber, *et al.,* (2002), demonstrate the idea that learning several types of NE simultaneously allows the finding of negative evidence (one type against all) and reduces over-generation. S. Cucerzan and Yarowsky (1999) also use a similar technique and apply it to many languages [12].

E. Riloff and Jones (1999) introduce mutual bootstrapping that consists of growing a set of entities and a set of contexts in turn [13]. Instead of working with predefined candidate NE's (found using a fixed syntactic construct), they start with a handful of seed entity examples of a given type (*e.g.,* Bolivia, Guatemala, Honduras are entities of type country) and accumulate all patterns found around these seeds in a large corpus. Contexts (e.g., offices in X, facilities in X…) are ranked and used to find new examples. Riloff and Jones note that the performance of that algorithm can deteriorate rapidly when noise is introduced in the entity list or pattern list. While they report relatively low precision and recall rate in their experiments, their work proved to be highly influential.

A. Cucchiarelli and Velardi (2001) use syntactic relations (*e.g.,* subject-object) to discover more accurate contextual evidence around the entities [14]. Again, this is a variant of E. Riloff and Jones mutual bootstrapping (1999). Interestingly, instead of using human generated seeds, they rely on existing NER systems (called early NE classifier) for initial NE examples.

M. Pasca, *et al.,* (2006) are also using techniques inspired by mutual bootstrapping [15]. However, they innovate through the use of D. Lin' s (1998) distributional similarity

to generate synonyms or, more generally, words which are members of the same semantic class allowing pattern generalization. For instance, for the pattern X was born in November, Lin's synonyms for November are {March, October, April, Mar, Aug., February, Jul, Nov., ...} thus allowing the induction of new patterns such as X was born in March. One of the contribution of Pasca, *et al.*, is to apply the technique to very large corpora (100 million web documents) and demonstrate that starting from a seed of 10 examples facts (defined as entities of type person paired with entities of type year - standing for the person year of birth) it is possible to generate one million facts with a precision of about 88%.

The problem of unlabeled data selection is addressed by J. Heng and Grishman (2006). They show how an existing NE classifier can be improved using bootstrapping methods. The main lesson they report is that relying upon large collection of documents is not sufficient by itself. Selection of documents using information retrieval-like relevance measures and selection of specific contexts that are rich in proper names and co-references bring the best results in their experiments.

3.3. Un-Supervised Learning

The typical approach in unsupervised learning is clustering. For example, one can try to gather named entities from clustered groups based on the similarity of context. There are other unsupervised methods too. Basically, the techniques rely on lexical resources (*e.g.*, WordNet), on lexical patterns and on statistics computed on a large un-annotated corpus. Here are some examples. E.Alfonseca and Manandhar (2002) study the problem of labeling an input word with an appropriate NE type [16]. NE types are taken from WordNet. The approach is to assign a topic signature to each WordNet synset by merely listing words that frequently co-occur with it in a large corpus. Then, given an input word in a given document, the word context (words appearing in a fixed-size window around the input word) is compared to type signature and classified under the most similar one.

In R. Evans (2003), the method for identification of hyponyms/hypernyms described in the work of M. Hearst (1992) is applied in order to identify potential hypernyms of sequences of capitalized words appearing in a document [17]. For instance, when X is a capitalized sequence, the query " such as X", is searched on the web and, in the retrieved documents, the noun that immediately precede the query can be chosen as the hypernym of X. Similarly, in P. Cimiano and Volker (2005), Hearst patterns are used. But this time, the feature consists of counting the number of occurrences of passages like: "city such as", "organization such as", *etc*.

Y. Shinyama and Sekine (2004) used an observation that named entities often appear synchronously in several news articles, whereas common nouns do not [18]. They found a strong correlation between being a named entity and appearing punctually (in time) and simultaneously in multiple news sources. This technique allows identifying rare named entities in an unsupervised manner and can be useful in combination with other NERC methods.

In O. Etzioni, *et al.*, (2005), Pointwise Mutual Information and Information Retrieval (PMI-IR) is used as a feature to assess that a named entity can be classified under a given type. PMI-IR, developed by P. Turney (2001), measures the dependence between two expressions using web queries [19]. A high PMI-IR means that expressions tend to co-occur. O. Etzioni, *et al.*, create features for each candidate entity (*e.g.*, London) and a large number of automatically generated discriminator phrases like "is a city", "nation of", *etc*.

4. Linguistic Issues and Challenges

4.1. Lack of Capitalization

Unlike languages like English that use the Latin script, where most NEs begin with a capital letter, capitalization is not a distinguishing orthographic feature of Uyghur script for recognizing person names. The ambiguity caused by the absence of this feature is further increased by the fact that most Uyghur person names (UPN) are indistinguishable from forms that are common nouns and adjectives (non-UPN). Thus, an approach relying only on looking up entries in UPN dictionaries would not be an appropriate way to tackle this problem, as ambiguous tokens/words that fall in this category are more likely to be used as non-UPN in text (Algahtani 2011). For example, the Uyghur proper name ئالىم (Alim) can be used in a sentence as a person name, a noun (Scientist). An UPN is usually found in a context, namely, with trigger and cue words to the left and/or right of the UPN. Therefore, it is common to resolve this type of ambiguity by analyzing the context surrounding the PN. However, this might require deeper analysis of the NE's context.

4.2. Agglutination

The agglutinative nature of Uyghur results in many different patterns that create many lexical variations. Each word may consist of a stem or root, and one or more suffixes in different combinations, resulting in a very systematic but complicated morphology. NER relies on the words forming the NE and the context in which it appears. Both the words and the contexts may appear in different inflected forms. In order to address data sparseness issues, it should be split with word segmentation. For example, the analysis of the word **قادىرنىڭ** (of Qadir) yields **قادىر** (Qadir) as a person name. Another solution is to omit all the affixes and keep only the root morpheme. This information is more convenient for NLP tasks that need to process these morphemes.

4.3. Ambiguity

Uyghur, like other languages, faces the problem of ambiguity between person names and non-person names. For example consider the following text:

يالقۇن بىلەن كۆرەش ياخشى ئاغىنىلەردىن ئىدى.

(Koresh is a good friend of Yalqun).

In this example, **كۆرەش** and **يالقۇن** is both a person name and proper adjective, thereby giving rise to a conflict situation. A boundary model was proposed by Jarulla Muhammad et al. (2014), and used boundary information to confirm the tag of words.

4.4. Lack of uniformity writing style

Uyghur has a high level of transcriptional ambiguity: A Person name can be transliterated in a multitude of ways. The lack of standardization is significant and leads to many variants of the same word that are spelled differently but still correspond to the same word with the same meaning, creating a many-to-one, ambiguity. For example, The Uyghur person name **مۇھەممەت** (Muhemmet) has another transcription **مەمەت** (Memet). One reason for this is that Uyghur has more speech sounds than Western European languages, which can ambiguously or erroneously lead to a Person name having more variants. One solution is to retain all versions of the name variants with a possibility of linking them together. Another solution is to normalize each occurrence of the variant; this requires a mechanism (such as string distance calculation) for name variant matching between a name variant and its normalized representation.

4.5. Lack of Resources

Large collections of tagged documents (corpora) as well as gazetteers (predefined lists of Person Name) are excellent sources that we can rely upon when implementing and testing the performance of a UPNR system. For making these linguistic resources to be useful, they should include unbiased distribution and representative that do not suffer from sparseness. Unfortunately, the available Uyghur resources for UPNR research often have limited capacity and coverage. Moreover, it is expensive to create or license these important Uyghur –language resources. For these reasons, researchers often rely on their own corpora, which require human annotation and verification. Few of these corpora have been made freely and publicly available for research purposes.

4.6. Person Name Recognition Tag Set

Tagging, also known as labeling, is the task of assigning a contextually appropriate tag (label) to every NE in the text. The sequence of words that is annotated with the same tag is considered a single multiword NE. The tag set used to tag NEs may differ according to user requirements.

Table 1. Tag Set for Person Name Recognition

Uyghur	English	Tags
ئەركىن	Erkin	B-PN
زتۇنيا	Tuniyaz	E-PN
بىلەن	with	O
ۋۇنسۇن	Tursun	B-PN
مەزنا	Governer	O
يىغىندا	in	conference
زۆس	a	speech
قىلدى	make	O

5. Recommendations

We believe the following recommendations will be helpful for the advancement of Uyghur person name recognition.

First of all, it is strongly recommended that Uyghur Named Entity evaluation data for training and testing has to be released. So, that people can use it freely and compare the performance of new methods. An open data source does provide a great forum of learning each other and prevent the redoing of some work.

Second, different methods should be comparatively tested instead of putting just one kind of method on experiment that only rule based approach or statistical approach. A good method in common for English NER does not necessarily will be the best one for all languages. Perhaps, some features will be quite important for English, but it will be useless for Uyghur. Some feature set with best accuracy may not give better result in Uyghur NER.

Thirdly, Uyghur Named Entity Recognition should be processed with integrated way that taking references from the approaches for all kind of Entity, such as person name, location name and organization name etc. Recent studies have been conducted separately to each NER type in Uyghur language. The integrated NER approach will be necessary for Uyghur NLP task in the future.

6. Conclusions

In this survey, we have shown Uyghur language Person Name Recognition Method and linguistic issues. More than twenty languages and a wide range of named entity types are studied. However, most of the work has concentrated on limited domains and textual genres such as news articles and web pages. We have also provided an overview of the techniques employed to develop UPNR systems. When supervised learning is used, a prerequisite is the availability of a large collection of annotated data. Such collections are available from the evaluation forums but remain rather rare and limited in domain and language coverage. Recent studies in the field have explored semi-supervised and unsupervised learning techniques that promise fast deployment for many entity types without the prerequisite of an annotated corpus. We have listed and categorized the features that are used in recognition and classification algorithms.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under grant of (61562081) and High Technology Research and Development Project of Xinjiang (201312103).

References

[1] D. Nadeau and S. Sekine "A survey of named entity recognition and classification", Lingvistica Investigations, vol. 30, no. 1, (**2007**), pp. 3-26.
[2] A. Rozi and Z. Chengqing, G. Mamateli and A. Hamdulla, "Approch to recognition Uyghur names based on conditional random fields", Journal of Tsinghua University(Sci & Tech), vol. 53, no. 6, **(2013)**, pp. 873-877.
[3] J. Muhammad and T. Ibrahim, "the approach based on statistics and rules", Journal of Xinjiang University (Natural Science Edition), no. 3, **(2014)**, pp. 319-324.
[4] D. M. Bikel and S. Miller, R. Schwartz, R. Weischedel, "Nymble: a High-Performance Learning Name-finder", In Proc. Conference on Applied Natural Language Processing, **(1997)**.
[5] S. Sekine, "Nyu: Description of the Japanese NE System Used For Met-2", In Proc. Message Understanding Conference, **(1998)**.
[6] B. Andrew, J. Sterling, E. Agichtein and R. Grishman, "NYU: Description of the MENE Named Entity System as used in MUC-7", In Proc. Seventh Message Understanding Conference, **(1998)**.
[7] A. Masayuki and Y. Matsumoto, "Japanese Named Entity Extraction with Redundant Morphological Analysis", In Proc. Human Language Technology conference - North American chapter of the Association for Computational Linguistics, **(2003)**.
[8] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons", In Proc. Conference on Computational Natural Language Learning, **(2003)**.
[9] N. David, P. Turney and S. Matwin, "Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity", In Proc. Canadian Conference on Artificial Intelligence, **(2006)**.
[10] Brin Sergey, "Extracting Patterns and Relations from the World Wide Web", In Proc. Conference of Extending Database Technology, Workshop on the Web and Databases, **(1998)**.
[11] C. Michael and Y. Singer, "Unsupervised Models for Named Entity Classification", In Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, **(1999)**.
[12] C. Silviu and D. Yarowsky, "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence", In Proc. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, **(1999)**.
[13] R. Ellen and R. Jones, "Learning Dictionaries for Information Extraction using Multi-level Bootstrapping", In Proc. National Conference on Artificial Intelligence, **(1999)**.
[14] C. Alessandro and P. Velardi, "Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence", Computational Linguistics, Cambridge: MIT Press, vol. 27, (**2001**), pp. 123-131.
[15] P. Marius, D. Lin, J. Bigham, A. Lifchits and A. Jain, "Organizing and Searching the World Wide Web of Facts—Step One: The One-Million Fact Extraction Challenge", In Proc. National Conference on Artificial Intelligence, **(2006)**.
[16] A. Enrique and S. Manandhar, "An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery", In Proc. International Conference on General WordNet, **(2002)**.

[17] E. Richard, "A Framework for Named Entity Recognition in the Open Domain", In Proc. Recent Advances in Natural Language Processing, **(2003)**.
[18] S. Yusuke and S. Sekine, "Named Entity Discovery Using Comparable News Articles", In Proc. International Conference on Computational Linguistics, **(2004)**.
[19] E. Oren, M. Cafarella, D. Downey, A. M., T. P. Shaked, S. Soderland and D. S. Weld and A. Yates, "Unsupervised Named-Entity Extraction from the Web: An Experimental Study", Artificial Intelligence, Essex: Elsevier Science Publishers, vol. 165, **(2005)**, pp. 91-134.

Authors

Tashpolat Nizamidin, He has received his B.E. degree in Electronics from Xinjiang University, China, in 2013. Currently, he is a M.S Student in Signal & Information processing in Xinjiang University; His research interest is Natural language Processing.

Palidan Tuerxun, She received her M. S. degree in 1996 from Liaoning University, China and her Ph.D. degree in 2015 from Northwestern University, China. Since 1992, she has been working as a teacher at Xinjiang University, and since 2004, she was an associate professor in school of software of Xinjiang University. Her research interests are machine learning and Uyghur natural language processing.

Askar Hamdulla, He received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 160 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.

Muhtar Arkin, He has received his B.S.in Electronic Information Science in June 2009 and M.S. in Signal and Information Processing in June 2013 at College of Information Science and Engineering, Xinjiang University, Urumqi, China. His research interest is Natural language processing.

