Name Tagging for Low-resource Incident Languages based on Expectation-driven Learning

Anonymous NAACL submission

Abstract

In this paper we tackle a challenging name tagging problem in an emergent setting - the tagger needs to be complete within a few hours for a new incident language (IL) using very few resources. Inspired by observing how human annotators attack this challenge, we propose a new expectation-driven learning framework. In this framework we rapidly acquire, categorize, structure and zoom in on IL-specific expectations (rules, features, patterns, gazetteers, etc.) from various non-traditional sources: consulting and encoding linguistic knowledge from native speakers, mining and projecting patterns from both mono-lingual and cross-lingual corpora, and typing based on cross-lingual entity linking. We also propose a cost-aware combination approach to compose expectations. Experiments on seven low-resource languages demonstrate the effectiveness and generality of this framework: we are able to setup a name tagger for a new IL within two hours, and achieve 33.8%-65.1% F-score.

1 Introduction: "Tibetan Room"

In many emergent situations such as disease outbreaks and natural disasters, there is great demand to rapidly develop a Natural Language Processing (NLP) system, such as name tagger, for a "surprise" Incident Language (IL) with very few resources. Traditional supervised learning methods that rely on large-scale manual annotations would be too costly.

Let's start by investigating how a human would discover information in a foreign IL environment. When we are in a foreign country, even if we don't know the language, we would still be able to guess the word "*gate*" from the airport broadcast based on its frequency and position in a sentence; guess the word "*station*" by pattern mining of many subway station labels; and guess the word "*left*" or "*right*" from a taxi driver's GPS speaker by matching movement actions. We designed a "***Tibetan Room***" game, similar to "*Chinese Room*" (Searle, 1980), by asking a human user who doesn't know Tibetan to find persons, locations and organizations from some Tibetan documents. We designed an interface where test sentences are presented to the player one by one. When the player clicks token, the interface will display up to 100 manually labeled Tibetan sentences that include this token. The player can also see translations of some common words and a small gazetteer of common names (800 entries) in the interface.

14 players who don't know Tibetan joined the game. Their name tagging F-scores ranged from 0% to 94%. We found that good players usually bring in some kind of "***expectations***" derived from their own native languages, or general linguistic knowledge, or background knowledge about the scenario. Then they actively search, confirm, adjust and update these expectations during tagging. For example, they know from English that location names are often ended with suffix words such as "*city*" and "*country*", so they search for phrases starting or ending with the translations of these suffix words. After they successfully tag some seeds, they will continue to discover more names based on more expectations. For example, if they already tagged an organization name *A*, and now observe a sequence matching a common English pattern "*[A (Organization)]'s [Title] [B (Person)]*", they will tag *B* as a person name. And if they know the scenario is about Ebola, they

will be looking for a phrase with translation similar to "*West Africa*" and tag it as a location. Similarly, based on the knowledge that names appear in a conjunction structure often have the same type, they propagate high-confidence types across multiple names. They also keep gathering and synthesizing common contextual patterns and rules (such as position, frequency and length information) about names and non-names to expand their expectations. For example, after observing a token frequently appearing between a subsidiary and a parent organization, they will predict it as a preposition similar to "*of*" in English, and tag the entire string as a nested organization.

Based on these lessons learned from this game, we propose to automatically acquire and encode expectations about what will appear in IL data (names, patterns, rules), and encode those expectations to drive IL name tagging. We explored various ways of systematically discovering and unifying latent and expressed expectations from nontraditional resources:

- **Language Universals**: Language-independent rules and patterns;
- **Native Speaker**: Interaction with native speakers through a machine-readable survey and supervised active learning;
- **Prior Mining**: IL entity prior knowledge mining from both mono-lingual and cross-lingual corpora and knowledge bases;

Furthermore, in emergent situations these expectations might not be available at once, and they may have different cost, so we need to organize and prioritize them to yield optimal performance within given time bounds. Therefore we also experimented with various **cost-aware** composition methods with the input of acquired expectations, plus a time bound for development (1 hour, 2 hours), and the output as a wall-time schedule that determines the best sequence of applying modules and maximizes the use of all available resources. Experiments on seven low-resource languages demonstrate that our framework can create an effective name tagger for an IL within a couple of hours using very few resources.

2 Starting Time: Language Universals

First we use some language universal rules, gazetteers and patterns to generate a binary feature vector $F = \{f_1, f_2, ...\}$ for each token. Table 1 shows these features along with examples. An identification rule is $r_I =< T_I, f = \{f_a, f_b, ...\} >$ where T_I is a "B/I/O" tag to indicate the beginning, inside or outside of a name, and $\{f_a, f_b, ...\}$ is a set of selected features. If the features are all matched, the token will be tagged as T_I. Similarly, a classification rule is $r_C =< T_C, f = \{f_a, f_b, ...\} >$, where T_C is "Person/Organization/Location". These rules are triggered in order, and some examples are as follows: <B, {AllUppercased}>, <PER, {PersonGaz}>, <ORG, {Capitalized, LongLength}> and etc.

3 Expectation Learning

3.1 Approach Overview

Figure 1 illustrates our overall approach of acquiring various expectations, by simulating the strategies human players adopted during the Tibetan Room game. Next we will present details about discovering expectations from each source.

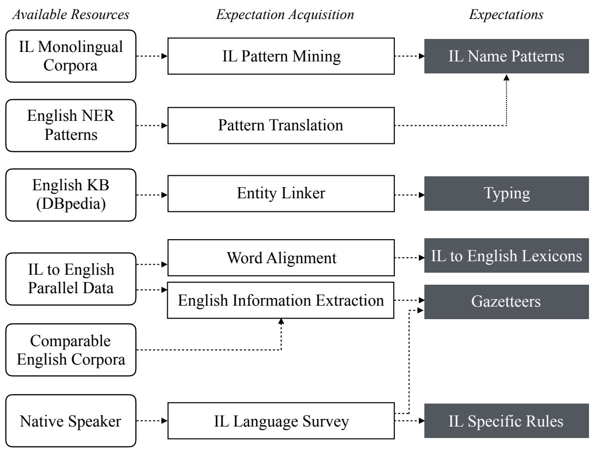


Figure 1: Expectation Driven Name Tagger Overview

3.2 Survey with Native Speaker

The best way to understand a language is to consult people who speak it. We introduce a human-in-the-loop process to acquire knowledge from native speakers. To meet the needs in the emergent setting, we design a comprehensive survey that aims to acquire a wide-range of IL-specific knowledge from native speakers in an efficient way. The survey categorizes questions and organizes them into a tree structure, so that the order of questions is chosen based on the answers of previous questions. The

Features	Examples (Feature name is underlined)
in English Gazetteer	- **PerGaz**: person $(472, 765)$; **LocGaz**: location $(211, 872)$; **OrgGaz**: organization $(124, 403)$; **Title** (889); **NoneName** $(2, 380)$.
Case	- **Capitalized**; - **AllUppercased**; - **MixedCase**
Punctuation	- **IternalPeriod**: includes an internal period
Digit	- **Digits**: consisted of digits
Length	- **LongLength**: a name including more than 4 tokens is likely to be an ORG
TF-IDF	- **TF-IDF**: if a capitalized word appears at the beginning of a sentence, and has a low TF-IDF, then it's unlikely to be a name
Patterns	- **Pattern1**: "*Title* $<$ PER Name $>$"
	- **Pattern2**: "$< PERName >, 00*$," where 00 are two digits
	- **Pattern3**: "$[< Name_i > ...], < Name_n - 1 >< singleterm >< Name_n >$" where all names have the same type.
Multi-occurrences	- **MultipleOccurrence**: If a word appears in both uppercased and lowercased forms in a single document, it's unlikely to be a name.

Table 1: Universal Name Tagger Features

survey answers are then automatically translated into rules, patterns or gazetteers in the tagger. Some example questions are shown in Table 2.

3.3 Mono-lingual Expectation Mining

We use a bootstrapping method to acquire IL patterns from unlabeled mono-lingual IL documents. Following the same idea in (Agichtein and Gravano, 2000; Collins and Singer, 1999), we first use names identified by high-confident rules as seeds, and generalize patterns from the contexts of these seeds. Then we evaluate the patterns and apply high-quality ones to find more names as new seeds. This process is repeated iteratively [1].

We define a pattern as a triple $< left, name, right >$[2], where $name$ is a name, left and right are context vectors with weighted terms. For example, from a Hausa sentence "***gwamnatin kasar Sin ta** samar wa kasashen yammacin Afirka ... (the Government of China has given ... products to the West African countries)*", we can discover a pattern $< \{< gwamnatin(government), 0.5 >, < kasar(country), 0.6 >\}, \{< Sin(China), 0.5 >\}, \{< ta(by), 0.2 >\} >$. This pattern matches strings like "*gwamnatin kasar Fiji by (by the government of Fiji)*".

For any two triples $t_i =< l_i, name_i, r_i >$ and $t_j =< l_j, name_j, r_j >$, we compute their similarity by:

$$Sim(t_i, t_j) = l_i \cdot l_j + r_i \cdot r_j$$

We use this similarity measurement to cluster all triples and select the centroid triples in each cluster as candidate patterns.

[1] We empirically set the number of iterations as 2 in this paper.

[2] Three tokens before and after

Similar to (Agichtein and Gravano, 2000), we evaluate the quality of a candidate pattern P by:

$$Conf(P) = \frac{P_{positive}}{(P_{positive} + P_{negative})}$$

,where $P_{positive}$ is the number of positive matches for P and $P_{negative}$ is the number of negative matches. Due to the lack of syntactic and semantic resources to refine these lexical patterns, we set a conservative confidence threshold 0.9.

3.4 Cross-lingual Expectation Projection

Name tagging research has been done for high-resource languages such as English for over twenty years, so we have learned a lot about them. We collected 1,362 patterns from English name tagging literature. Some examples are listed below:

- $< \{\}, \{PER\}, \{< say >, < . >\} >$
- $< \{< headquarter >, < in >\}, \{LOC\}, \{\} >$
- $< \{< secretary >, < of >\}, \{ORG\}, \{\} >$
- $< \{< in >, < the >\}, \{LOC\}, \{< area >\} >$

Besides the static knowledge like patterns, we can also dynamically acquire expected names from topically-related English documents for a given IL document. We apply the Stanford name tagger (Finkel et al., 2005) to the English documents to obtain a list of expected names. Then we translate the English patterns and expected names to IL. When there is no human constructed English-to-IL lexicon available, we derive a word-for-word translation table from a small parallel data set using the GIZA++ word alignment tool (Och and Ney, 2003). We also convert IL text to Latin characters based on Unicode mapping[3], and then apply Soundex code (Mortimer and Salathiel, 1995; Raghavan and Allan, 2004) to

[3] http://www.ssec.wisc.edu/ tomw/java/unicode.html

True/False Questions
1. The letters of this language have upper and lower cases
2. The names of people, organizations and locations start with a capitalized (uppercased) letter
3. The first word of a sentence starts with a capitalized (uppercased) letter
4. Some periods indicate name abbreviations, e.g., St. = Saint, I.B.M. = International Business Machines.
5. Locations usually include designators, e.g., in a format like "country United states", "city Washington"
6. Some prepositions are part of names
Text input
1. Morphology: please enter preposition suffixes as many as you can (e.g. "'da" in "Ankara'da yaşıyorum (I live in Ankara)" is a preposition suffix which means "in").
Translation
1. Please translate the following English words and phrases:
- organization suffix: agency, group, council, party, school, hospital, company, office, ...
- time expression: January, ..., December; Monday, ..., Sunday; ...

Table 2: Survey Question Examples

find the IL name equivalent that shares the most similar pronunciation as each English name. For example, the Bengali name "টনি ব্লেয়ার" and "*Tony Blair*" have the same Soundex code "*T500 B460*".

3.5 Mining Expectations from KB

In addition to unstructured documents, we also try to leverage structured English knowledge bases (KBs) such as DBpedia[4]. Each entry is associated with a set of types such as `Company`, `Actor` and `Agent`. We utilize the Abstract Meaning Representation corpus (Banarescu et al., 2013) which contains both entity type and linked KB title annotations, to automatically map $9,514$ entity types in DBPedia to three main entity types of interest: Person (PER), Location (LOC) and Organization (ORG).

Then we adopt a language-independent cross-lingual entity linking system (Wang et al., 2015) to link each IL name mention to English DBPedia. This linker is based on an unsupervised quantified collective inference approach. It constructs knowledge networks from the IL source documents based on entity mention co-occurrence, and knowledge networks from KB. Each IL name is matched with candidate entities in English KB using name translation pairs derived from inter-lingual KB links in Wikipedia and DBPedia. We also apply the word-for-word translation tables constructed from parallel data as described in Section 3.4 to translate some uncommon names. Then it performs semantic comparison between two knowledge networks based on three criteria: salience, similarity and coherence. Finally we map the DBPedia types associated with the linked entity candidates to obtain the entity type for each IL name.

[4]http://dbpedia.org

4 Supervised Active Learning

We anticipated that not all expectations can be encoded as explicit rules and patterns, or covered by projected names, therefore for comparison we introduce a supervised method with pool-based active learning to learn implicit expectations (features, new names, etc.) directly from human data annotation. We exploited basic lexical features including ngrams, adjacent tokens, casing information, punctuations and frequency to train a Conditional Random Fields (CRFs) (Lafferty et al., 2001) based on model through active learning.

We segment documents into sentences and use each sentence as a training unit. Let $\mathbf{x}_b^*$ be the most informative instance according to a query strategy $\phi(\mathbf{x})$, which is a function used to evaluate each instance $\mathbf{x}$ in the unlabeled pool U. Algorithm 1 illustrates the procedure.

Algorithm 1 Pool-based Active Learning

```
L ← labeled set, U ← unlabeled pool
φ(·) ← query strategy, B ← query batch size
M ← maximum number of tokens
while Length(L) < M do
    θ = train(L);
    for b ∈ {1, 2, ..., B} do
        x*_b = arg max_{x∈U} φ(x)
        L = L ∪ {x*_b, label(x*_b)}
        U = U − x*_b
    end for
end while
```

We use ϕ^{SE} to represent how informative a sentence is, defined as **sentence entropy (SE)**:

$$\phi^{SE}(\mathbf{x}) = -\sum_{t=1}^{T}\sum_{m=1}^{M} P_\theta(y_t = m) log P_\theta(y_t = m)$$

, where T is the length of **x**, m ranges over all possible token labels and $P_\theta(y_t = m)$ is the probability when y_t is tagged as m.

5 Cost-aware Combination

A new requirement for IL name tagging is a **Linguistic Workflow Generator**, which can generate an activity schedule to organize and maximize the use of acquired expectations to yield optimal F-scores within given time bounds. Therefore, the input to the IL name tagger is not only the test data, but also a time bound for development (1 hour, 2 hours, 24 hours, 1 week, 1 month, etc.).

Figure 2 illustrates our cost-aware expectation composition approach. Given some IL documents as input, as the clock ticks, the system delivers name tagging results at time 0 (immediately), time 1 (e.g., in one hour) and time 2 (e.g., in two hours). At time 0, name tagging results are provided by the universal tagger described in Section 2. During the first hour, we can either ask the native speaker to annotate a small amount of data for supervised active learning of a CRFs model, or fill in the survey to build a rule-based tagger. We estimated the confidence value of each expectation-driven rule based on a small development set. When the results of two taggers are conflicting, if the applied rule has high confidence we will trust its output, otherwise adopt the CRFs model's output.

6 Experiments

In this section we will present our experimental details, results and observations.

6.1 Data

We evaluate our framework on seven low-resource incident languages: Bengali, Hausa, Tagalog, Tamil, Thai, Turkish and Yoruba, using the ground-truth name tagging annotations from the DARPA LORELEI program [5]. Table 3 shows data statistics.

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

Language	IL Test Docs	Name	Unique Name	IL Dev. Docs	IL-English Docs
Bengali	100	4,713	2,820	12,495	169
Hausa	100	1,619	950	13,652	645
Tagalog	100	6,119	3,375	1,616	145
Tamil	100	4120	2,871	4,597	166
Thai	100	4,954	3,314	10,000	191
Turkish	100	2,694	1,323	10,000	484
Yoruba	100	3,745	2,337	427	252

Table 3: Data Statistics

6.2 Cost-aware Overall Performance

We test with three checking points: starting time, within one hour, and within two hours. Based on the combination approach described in Section 5, we can have three possible combinations of the expectation-driven learning and supervised active learning methods during two hours: (1) expectation-driven learning + supervised active learning; (2) supervised active learning + expectation-driven learning; and (3) supervised active learning for two hours. Figure 4 compares the overall performance of these combinations for each language.

We can see that our approach is able to rapidly set up a name tagger for an IL and achieves promising performance. During the first hour, there is no clear winner between expectation-driven learning or supervised active learning. But it's clear that supervised active learning for two hours is generally not the optimal solution. Using Hausa as a case study, we take a closer look at the supervised active learning curve as shown in Figure 3. We can see that supervised active learning based on simple lexical features tends to converge quickly. As time goes by it will reach its own upper-bound of learning and generalizing linguistic features. In these cases our proposed expectation-driven learning method can compensate by providing more explicit and deeper IL-specific linguistic knowledge.

6.3 Comparison of Expectation Discovery Methods

Table 4 shows the performance gain of each type of expectation acquisition method. IL gazetteers covered some common names, especially when the universal case-based rules failed at identifying names from non-Latin languages. IL name patterns were mainly effective for classification. For example, the Tamil name "கத்தோலிக்கன் சிரியன் வங்கியில (Catholic Syrian Bank)" was classi-

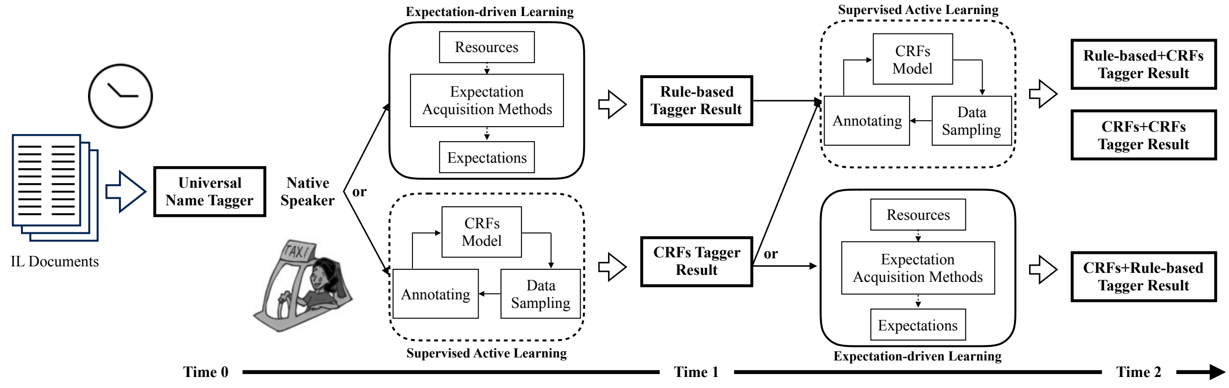


Figure 2: Cost-aware Expectation Composition

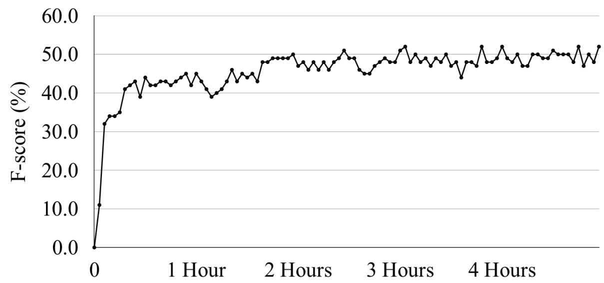


Figure 3: Hausa Supervised Active Learning Curve

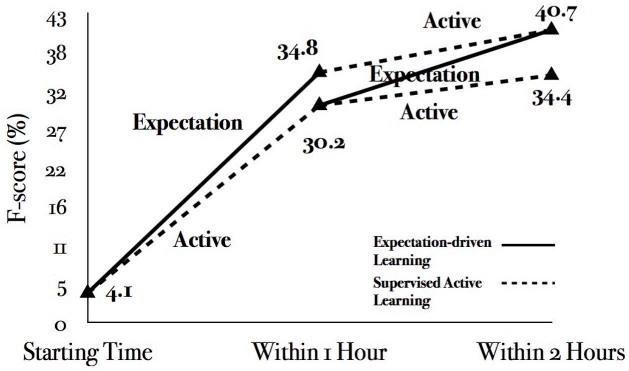
fied as an organization because it ends with an organization suffix word "வங்கியில (bank)". The patterns projected from English were proven very effective at identifying name boundaries. For example, some non-names such as titles are also capitalized in Turkish, so simple case-based patterns produced many spurious names. But projected patterns can fix many of them. In the following Turkish sentence, "*Ancak Avrupa Birliği Dış İlişkiler Sorumlusu Catherine Ashton,...(But European Union foreign policy chief Catherine Ashton,...)*", among all these capitalized tokens, after we confirmed "*Avrupa Birliği (European Union)*" as an organization and "*Dış İlişkiler Sorumlusu (foreign policy chief)*" as a title, we applied a pattern projected from English "*[Organization] [Title] [Person]*" and successfully identified "*Catherine Ashton*" as a person. Cross-lingual entity linking based typing successfully enhanced classification accuracy, especially for languages where names often appear the same as their English forms and so entity linking achieved high accuracy. For example, "*George Bush*" keeps the same in Hausa, Tagalog and Yoruba as English.

6.4 Impact of Supervised Active Learning

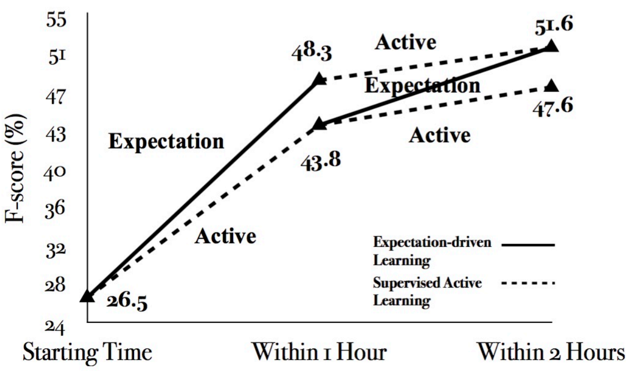
Figure 5 shows the comparison of supervised active learning and passive learning (random sampling in training data selection). We asked a native speaker to annotate Chinese news documents in one hour, and estimated the human annotation speed approximately as 7,000 tokens per hour. Therefore we set the number of tokens as 7,000 for one hour, and 14,000 for two hours. We can clearly see that supervised active learning significantly outperforms passive learning for all languages, especially for Tamil, Tagalog and Yoruba. Because of the rich morphology in Turkish, the gain of supervised active learning is relatively small because simple lexical features cannot capture name-specific characteristics regardless of the size of labeled data. For example, some prepositions (e.g., "*nin (in)*") can be part of the names, so it's difficult to determine name boundaries, such as "*<ORG Ludian bölgesi **hastanesi**>**nin** (in <ORG Ludian Hospital>)*"

6.5 Remaining Error Analysis

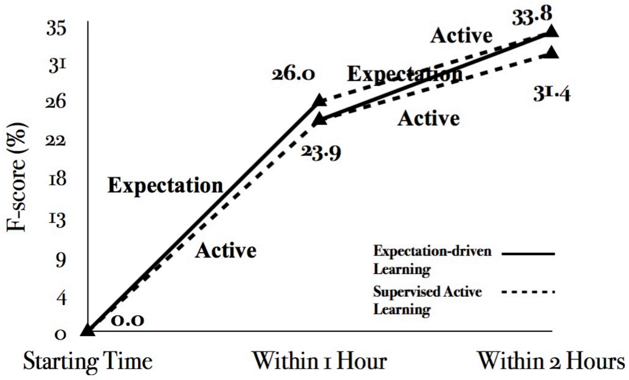
Table 5 presents the detailed break-down scores for all languages. We can see that name identification, especially organization identification is the main bottleneck for all languages. For example, many organization names in Hausa are often very long, nested or all low-cased, such as "*makarantar horas da Malaman makaranta ta Bawa Jan Gwarzo (Bawa Jan Gwarzo Memorial Teachers College)*" and "*kungiyar masana'antu da tattalin arziki ta kasar Sin (China's Association of Business and Industry)*".



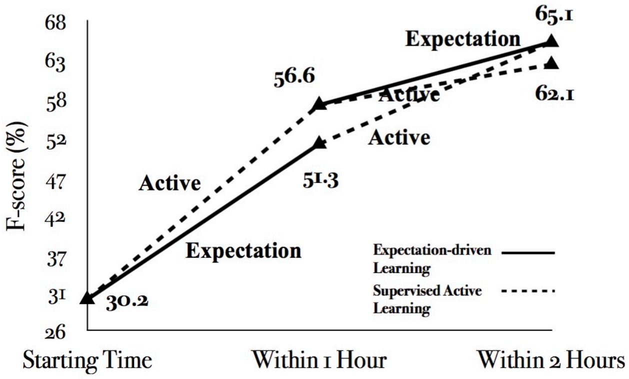
(a) Bengali



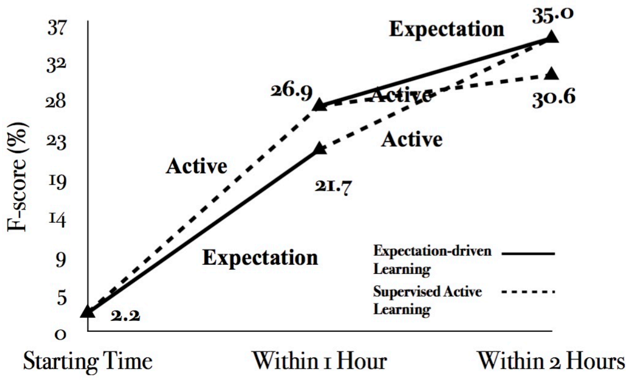
(b) Hausa



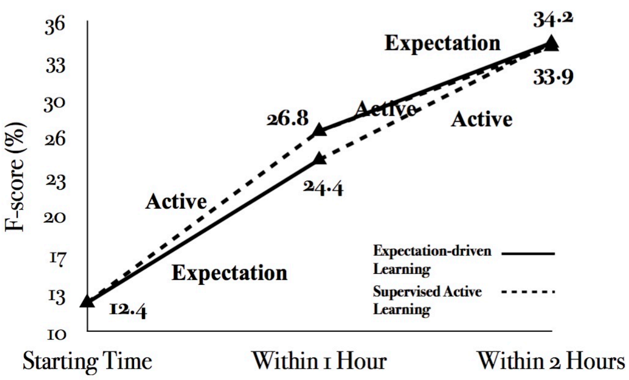
(c) Tamil



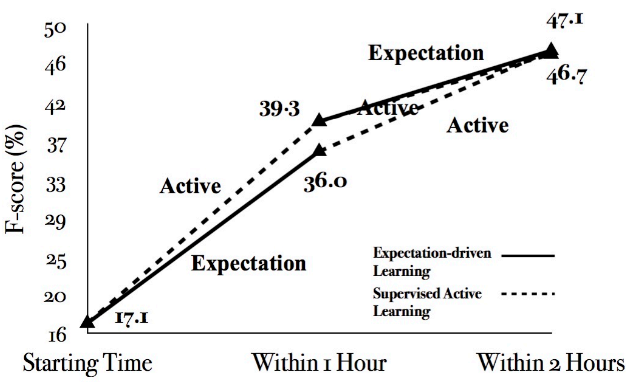
(d) Tagalog



(e) Thai



(f) Turkish



(g) Yoruba

Figure 4: Comparison of methods combining expectation-driven learning and supervised active learning given various time bounds

Methods	Bengali	Hausa	Tamil	Tagalog	Thai	Turkish	Yoruba
Universal Rules	4.1	26.5	0.0	30.2	2.2	12.4	17.1
+IL Gazetteers	29.7	32.1	21.8	34.3	18.9	17.3	26.9
+IL Name Patterns	31.2	33.8	22.9	35.1	18.9	19.1	28.0
+IL to English Lexicons	31.3	35.2	24.0	38.0	20.5	19.6	29.4
+KB Linking based Typing	34.0	44.8	25.1	51.1	20.7	24.2	35.1
+IL Rules and Gazetteers	34.8	48.3	26.0	51.3	21.7	24.4	36.0

Table 4: Contributions of Various Expectation Discovery Methods (F-score %)

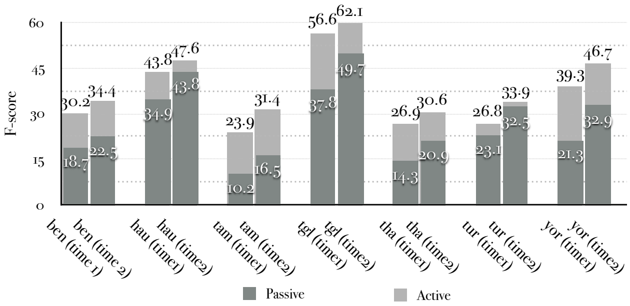


Figure 5: Active Learning vs. Passive Learning (%)

Language	Identification F-score				Typing Accuracy	Overall F-score
	PER	ORG	LOC	All		
Bengali	51.0	32.7	54.3	48.5	84.1	40.7
Hausa	51.8	36.6	63.3	55.1	93.6	51.6
Tamil	40.4	16.4	46.8	39.2	86.2	33.8
Tagalog	71.6	65.2	73.9	70.1	92.8	65.1
Thai	48.5	21.8	72.8	48.6	72.0	35.0
Turkish	59.3	36.7	31.0	40.7	84.1	34.2
Yoruba	69.3	38.3	60.0	57.2	82.3	47.1

Table 5: Breakdown Scores

7 Related Work

Name Tagging is a well-studied problem. Many types of frameworks have been used, including rules (Nadeau and Sekine, 2007; Farmakiotou et al., 2000), supervised models using monolingual labeled data (McCallum and Li, 2003; Zhou and Su, 2002; Chieu and Ng, 2002; Rizzo and Troncy, 2012), bilingual labeled data (Li et al., 2012; Kim et al., 2012; Che et al., 2013; Wang et al., 2013) or naturally partially annotated data such as Wikipedia (Nothman et al., 2013), bootstrapping (Chiticariu et al., 2010; Wu et al., 2009; Niu et al., 2003; Agichtein and Gravano, 2000; Becker et al., 2005), and unsupervised learning (Mikheev et al., 1999; McCallum and Li, 2003; Etzioni et al., 2005; Nadeau et al., 2006; Nadeau and Sekine, 2007; Ji and Lin, 2009). It's been explored for many non-English languages such as in Chinese (Ji and Grishman, 2005; Li et al., 2014), Japanese (Asahara and Matsumoto, 2003; Li et al., 2014), Arabic (Maloney and Niv, 1998), Catalan (Carreras et al., 2003), Bulgarian (Osenova and Kolkovska, 2002), Dutch (De Meulder et al., 2002), French (Béchet et al., 2000), German (Thielen, 1995), Italian (Cucchiarelli et al., 1998), Greek (Karkaletsis et al., 1999), Spanish (Arévalo et al., 2002), Portuguese (Hana et al., 2006), Serbo-croatian (Nenadić and Spasić, 2000), Swedish (Dalianis and Åström, 2001) and Turkish (Tür et al., 2003). However, most of previous work relied on substantial amount of resources such as language-specific rules, basic tools such as part-of-speech taggers, a large amount of labeled data, or a huge amount of Web ngram data, which are usually unavailable for low-resource ILs. In contrast, in this paper we put the name tagging task in a new emergent setting where we need to process a surprise IL within very short time using very few resources. The results of the tested ILs are still far from perfect, but we hope our detailed comparison and result analysis can introduce new ideas to balance the quality and cost of name tagging.

8 Conclusions and Future Work

Name tagging for a new IL is a very important but also challenging task. We conducted a thorough study on various ways of acquiring, encoding and composing expectations from multiple non-traditional sources. Experiments demonstrate that this framework can be used to build a promising name tagger for a new IL within a few hours. In the future we will exploit broader and deeper entity prior knowledge to improve name identification. We will aim to make the framework more transparent for native speakers so the survey can be done in an automatic interactive question-answering fashion. We will also develop methods to make the tagger capable of active self-assessment to produce the best workflow within time bounds.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*.

Montse Arévalo, Xavier Carreras, Lluís Màrquez, María Antònia Martí, Lluís Padró, and María José Simón. 2002. A proposal for wide-coverage spanish named entity recognition. *Procesamiento del lenguaje natural*.

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

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*.

Frédéric Béchet, Alexis Nasr, and Franck Genet. 2000. Tagging unknown proper names using decision trees. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*.

Markus Becker, Ben Hachey, Beatrice Alex, and Claire Grover. 2005. Optimising selective sampling for bootstrapping named entity recognition. In *Proceedings of ICML-2005 Workshop on Learning with Multiple Views*.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. Named entity recognition for catalan using spanish resources. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of HLT-NAACL*.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics*.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Alessandro Cucchiarelli, Danilo Luzi, and Paola Velardi. 1998. Automatic semantic tagging of unknown proper names. In *Proceedings of the 17th international conference on Computational linguistics*.

Hercules Dalianis and Erik Åström. 2001. Swenam—a swedish named entity recognizer. Technical report, Technical Report. Department of Numerical Analysis and Computing Science.

Fien De Meulder, Walter Daelemans, and Véronique Hoste. 2002. A named entity recognition system for dutch. *Language and Computers*.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*.

Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

Jirka Hana, Anna Feldman, Chris Brew, and Luiz Amaral. 2006. Tagging portuguese with a spanish tagger using cognates. In *Proceedings of the International Workshop on Cross-Language Knowledge Induction*.

Heng Ji and Ralph Grishman. 2005. Improving name tagging by reference resolution and relation detection. In *Proceedings of ACL2005*.

Heng Ji and Dekang Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of PACLIC2009*.

Vangelis Karkaletsis, Georgios Paliouras, Georgios Petasis, Natasa Manousopoulou, and Constantine D Spyropoulos. 1999. Named-entity recognition from greek and english texts. *Journal of Intelligent and Robotic Systems*.

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*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM international conference on Information and knowledge management*.

Haibo Li, Masato Hagiwara, Qi Li, and Heng Ji. 2014. Comparison of the impact of word segmentation on name tagging for chinese and japanese. In *Proceedings of LREC2014*.

John Maloney and Michael Niv. 1998. Tagarab: a fast, accurate arabic name recognizer using high-precision morphological analysis. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*.

JY Mortimer and JA Salathiel. 1995. 'soundex'codes of surnames provide confidentiality and accuracy in a national hiv database. *Communicable disease report. CDR review*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*.

David Nadeau, Peter Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity.

Goran Nenadić and Irena Spasić. 2000. Recognition and acquisition of compound names from corpora. In *Natural Language Processing—NLP 2000*.

Cheng Niu, Wei Li, Jihong Ding, and Rohini K Srihari. 2003. Bootstrapping for named entity tagging using concept-based seeds. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. 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*.

Petya Osenova and Sia Kolkovska. 2002. Combining the named-entity recognition task and np chunking strategy for robust pre-processing. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September*.

Hema Raghavan and James Allan. 2004. Using soundex codes for indexing names in asr documents. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*.

Giuseppe Rizzo and Raphaël Troncy. 2012. Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

John Searle. 1980. Minds, brains, and programs. *Journal of the Association for Computing Machinery*.

Christine Thielen. 1995. An approach to proper name tagging for german. *arXiv preprint cmp-lg/9506024*.

Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction system for turkish. *Natural Language Engineering*.

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*.

Han Wang, Jin Guang Zheng, Xiaogang Ma, Peter Fox, and Heng Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. 2009. Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.