ELISA-EDL: A Cross-lingual Entity Extraction, Linking and Localization System

Boliang Zhang[1], **Ying Lin**[1], **Xiaoman Pan**[1], **Di Lu**[1], **Jonathan May**[2], **Kevin Knight**[2], **Heng Ji**[1]
[1] Rensselaer Polytechnic Institute
{zhangb8,liny9,panx2,lud2,jih}@rpi.edu
[2] Information Sciences Institute
{jonmay,knight}@isi.edu

Abstract

We demonstrate ELISA-EDL, a state-of-the-art re-trainable system to extract entity mentions from low-resource languages, link them to external English knowledge bases, and visualize locations related to disaster topics on a world heatmap. We make all of our data sets[1], resources and system training and testing APIs[2] publicly available for research purpose.

1 Introduction

Our cross-lingual entity extraction, linking and localization system is capable of extracting named entities from unstructured text in any of 282 Wikipedia languages, translating them into English, and linking them to English Knowledge Bases (Wikipedia and Geonames). This system then produces visualizations of the results such as heatmaps, and thus it can be used by an English speaker for monitoring disasters and coordinating rescue and recovery efforts reported from incident regions in low-resource languages. In the rest of the paper, we will present a comprehensive overview of the system components (Section 2 and Section 3), APIs (Section 4), interface[3](Section 5), and visualization[4] (Section 6).

2 Entity Extraction

Given a text document as input, the entity extraction component identifies entity name mentions and classifies them into pre-defined types: Person (PER), Geo-political Entity (GPE), Organization (ORG) and Location (LOC). We consider name tagging as a sequence labeling problem, to tag each token in a sentence as the Beginning (B), Inside (I) or Outside (O) of an entity mention with a certain type. Our model is based on a bi-directional long short-term memory (LSTM) networks with a Conditional Random Fields (CRFs) layer (Chiu and Nichols, 2016). It is challenging to perform entity extraction across a massive variety of languages because most languages don't have sufficient data to train a machine learning model. To tackle the low-resource challenge, we developed creative methods of deriving noisy training data from Wikipedia (Pan et al., 2017), exploiting non-traditional language-universal resources (Zhang et al., 2016) and cross-lingual transfer learning (Cheung et al., 2017).

3 Entity Linking and Localization

After we extract entity mentions, we link GPE and LOC mentions to GeoNames[5], and PER and ORG mentions to Wikipedia[6]. We adopt the name translation approach described in (Pan et al., 2017) to translate each tagged entity mention into English, then we apply an unsupervised collective inference approach (Pan et al., 2015) to link each translated mention to the target KB. Figure 2 shows an example output of a Hausa document. The extracted entity mentions "*Stephane Dujarric*" and "*birnin Bentiu*" are linked to their corresponding entries in Wikipedia and GeoNames respectively.

Compared to traditional entity linking, the unique challenge of linking to GeoNames is that it is very scarce, without rich linked structures or text descriptions. Only 500k out of 4.7 million entities in Wikipedia are linked to GeoNames. Therefore, we associate mentions with entities in the KBs in a collective manner, based on salience, similarity and coherence measures (Pan et al., 2015). We calculate topic-sensitive PageRank scores for 500k overlapping entities between

[1] https://elisa-ie.github.io/wikiann
[2] https://elisa-ie.github.io/api
[3] https://elisa-ie.github.io
[4] https://elisa-ie.github.io/heatmap
[5] http://www.geonames.org
[6] https://www.wikipedia.org

Proceedings of NAACL-HLT 2018: Demonstrations, pages 41–45
New Orleans, Louisiana, June 2 - 4, 2018.

APIs	Description
/status	Retrieve the current server status, including supported languages, language identifiers, and the state (offline, online, or pending) of each model.
/status/{identifier}	Retrieve the current status of a given language.
/entity_discovery_and_linking/{identifier}	Main entry of the EDL system. Take input in either plain text or *.ltf format, tag names that are PER, ORG or LOC/GPE, and link them to Wikipedia.
/name_transliteration/{identifier}	Transliterate a name to Latin script.
/entity_linking/{identifier}	Query based entity linking. Link each mention to KBs.
/entity_linking_amr	English entity linking for Abstract Meaning Representation (AMR) style input (Pan et al., 2015). AMR (Banarescu et al., 2013) is a structured semantic representation scheme. The rich semantic knowledge in AMR boosts linking performance.
/localize/{identifier}	Localize a LOC/GPE name based on GeoNames database.

Table 1: RUN APIs description.

APIs	Description
/status	An alias of /status
/status/{identifier}	Query the current status of a model being trained.
/train/{identifier}	Train a new name tagging model for a language. A model id is automatically generated and returned based on model name, and time stamp.

Table 2: TRAIN APIs description.

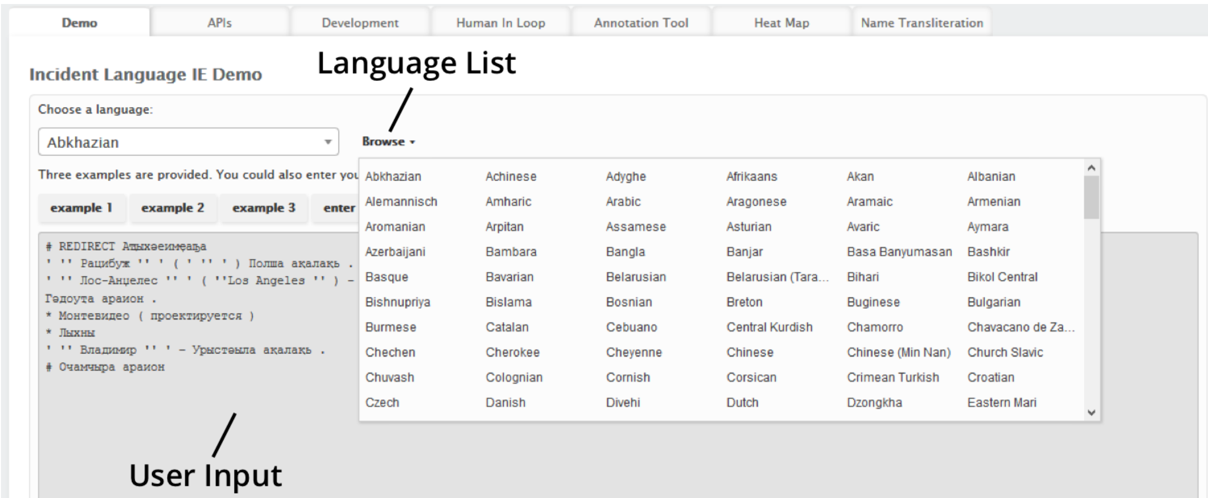

Figure 1: Cross-lingual Entity Extraction and Linking Interface

Figure 2: Cross-lingual Entity Extraction and Linking Testing Result Visualization

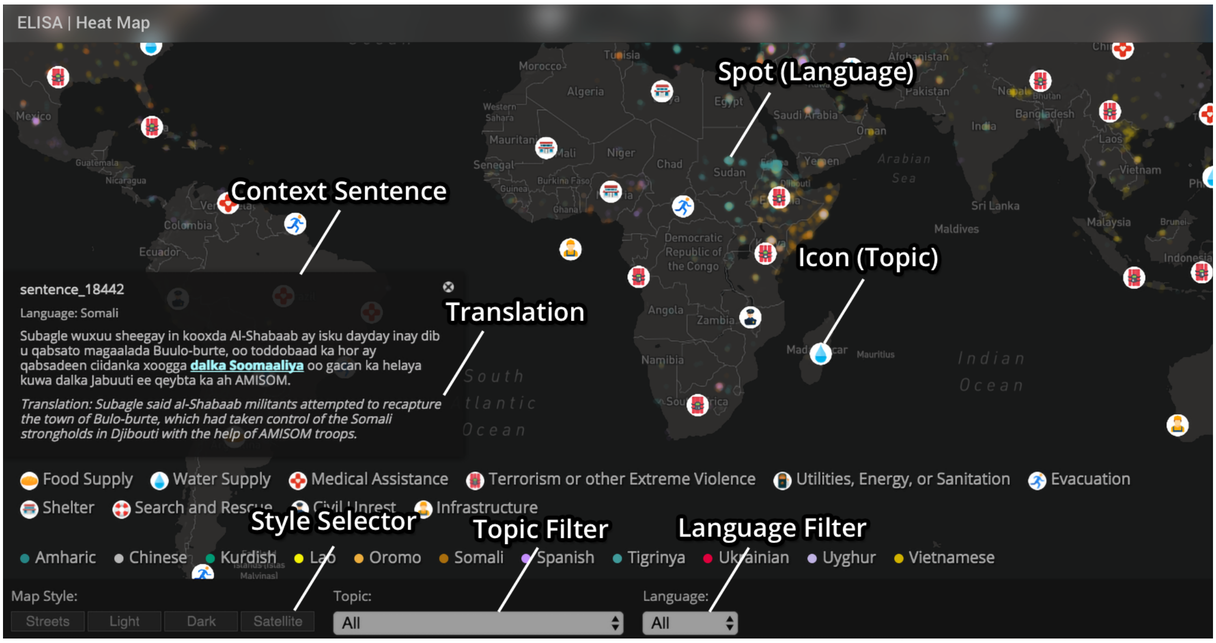

Figure 3: Heatmap Visualization

Language	F1 (%)	Language	F1 (%)
Arabic	51.9	Bengali	74.8
Chechen	58.9	Persian	58.4
Hausa	70.2	Hungarian	60.2
Oromo	81.3	Russian	63.7
Somali	67.6	Tamil	65.9
Thai	69.8	Tigrinya	73.2
Tagalog	78.7	Turkish	74.4
Uyghur	72.3	Uzbek	71.8
Vietnamese	68.5	Yoruba	50.1

Table 3: Name Tagging Performance on Low-Resource Languages

GeoNames and Wikipedia as their salience scores. Then we construct knowledge networks from source language texts, where each node represents a entity mention, and each link represents a sentence-level co-occurrence relation. If two mentions cooccur in the same sentence, we prefer their entity candidates in the GeoNames to share an administrative code and type, or be geographically close in the world, as measured in terms of latitude and longitude.

Table 3 shows the performance of our system on some representative low-resource languages for which we have ground-truth annotations from the DARPA LORELEI[7] programs, prepared by the Linguistic Data Consortium.

4 Training and Testing APIs

In this section, we introduce our back-end APIs. The back-end is a set of RESTful APIs built with Python Flask[8], which is a light weight framework that includes template rendering and server hosting capabilities. We use Swagger for documentation management. Besides the on-line hosted APIs, we also publish our Docker copy[9] at Dockerhub for software distribution.

In general, we categorize the APIs into two sections: RUN and TRAIN. The RUN section is responsible for running the pre-trained models for 282 languages, and the TRAIN section provides a re-training function for users who want to train their own customized name tagging models using their own datasets. We also published our training and test data sets, as well as resources related to at morphology analysis and name translation at: `https://elisa-ie.github.io/wikiann`. Table 1 and Table 2 present the detailed functionality and usages of the APIs of these two sections. Besides the core components as described in Section 2 and Section 3, we also provide the APIs of additional components, including a re-trainable name transliteration component (Lin et al., 2016) and a universal name and word translation component based on word alignment derived from cross-

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

[8] `http://flask.pocoo.org`

[9] `https://hub.docker.com/r/elisarpi/elisa-ie/`

lingual Wikipedia links (Pan et al., 2017). More detailed usages and examples can be found in our Swagger[10] documentation: `https://elisa-ie.github.io/api`.

5 Testing Interface

Figure 1 shows the test interface, where a user can select one of the 282 languages, enter a text or select an example document, and run the system. Figure 2 shows an output example. In addition to the entity extraction and linking results, we also display the top 5 images for each entity retrieved from Google Image Search[11]. In this way even when a user cannot read a document in a low-resource language, s/he will obtain a high-level summary of entities involved in the document.

6 Heatmap Visualization

Using disaster monitoring as a use case, we detect the following ten topics from the input multilingual data based on translating 117 English disaster keywords via PanLex[12]: (1) water supply, (2) food supply, (3) medical assistance, (4) terrorism or other extreme violence, (5) utilities, energy or sanitation, (6) evacuation, (7) shelter, (8) search and rescue, (9) civil unrest or widespread crime, and (10) infrastructure, as defined in the NIST LoreHLT2017 Situation Frame detection task[13]. If a sentence includes one of these topics and also a location or geo-political entity, we will visualize the entity on a world *heatmap* using Mapbox[14] based on its coordinates in the GeoNames database obtained from the entity linker. We also show the entire context sentence and its English translation produced from our state-of-the-art Machine Translation system for low-resource languages (Cheung et al., 2017). Figure 3 illustrates an example of the visualized heatmap.

We use different colors and icons to stand for different languages and frame topics respectively (e.g., the bread icon represents "food supply"). Users can also specify the language or frame topic or both to filter out irrelevant results on the map. By clicking an icon, its context sentence will be displayed in a pop-up with automatic translation and highlighted mentions and keywords. We provide various map styles (light, dark, satellite, and streets) for different needs, as shown in Figure 4.

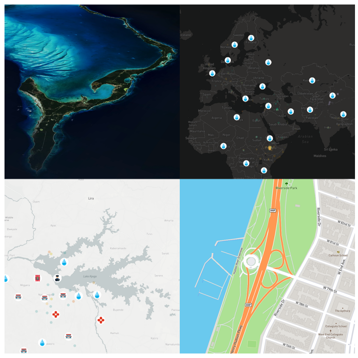

Figure 4: Different Map Styles

7 Related Work

Some recent work has also focused on low-resource name tagging (Tsai et al., 2016; Littell et al., 2016; Zhang et al., 2016; Yang et al., 2017) and cross-lingual entity linking (McNamee et al., 2011; Spitkovsky and Chang, 2011; Sil and Florian, 2016), but the system demonstrated in this paper is the first publicly available end-to-end system to perform both tasks and all of the 282 Wikipedia languages.

8 Conclusions and Future Work

Our publicly available cross-lingual entity extraction, linking and localization system allows an English speaker to gather information related to entities from 282 Wikipedia languages. In the future we will apply common semantic space construction techniques to transfer knowledge and resources from these Wikipedia languages to all thousands of living languages. We also plan to significantly expand entities to the thousands of fine-grained types defined in YAGO (Suchanek et al., 2007) and WordNet (Miller, 1995).

Acknowledgments

This work was supported by the U.S. DARPA LORELEI Program No. HR0011-15-C-0115. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

[10] `https://swagger.io`
[11] `https://images.google.com`
[12] `http://panlex.org`
[13] `https://www.nist.gov/itl/iad/mig/lorehlt-evaluations`
[14] `https://www.mapbox.com`

References

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*.

Leon Cheung, Thamme Gowda, Ulf Hermjakob, Nelson Liu, Jonathan May, Alexandra Mayn, Nima Pourdamghani, Michael Pust, Kevin Knight, Nikolaos Malandrakis, Pavlos Papadopoulos, Anil Ramakrishna, Karan Singla, Victor Martinez, Colin Vaz, Dogan Can, Shrikanth Narayanan, Kenton Murray, Toan Nguyen, David Chiang, Xiaoman Pan, Boliang Zhang, Ying Lin, Di Lu, Lifu Huang, Kevin Blissett, Tongtao Zhang, Heng Ji, Ondrej Glembek, Murali Karthick Baskar, Santosh Kesiraju, Lukas Burget, Karel Benes, Igor Szoke, Karel Vesely, Jan "Honza" Cernocky, Camille Goudeseune, Mark Hasegawa Johnson, Leda Sari, Wenda Chen, and Angli Liu. 2017. ELISA system description for lorehlt 2017. In *Proc. LoReHLT2017*.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4:357–370.

Ying Lin, Xiaoman Pan, Aliya Deri, Heng Ji, and Kevin Knight. 2016. Leveraging entity linking and related language projection to improve name transliteration. In *Proc. ACL2016 Workshop on Named Entities*.

Patrick Littell, Kartik Goyal, David Mortensen, Alexa Little, Chris Dyer, and Lori Levin. 2016. Named entity recognition for linguistic rapid response in low-resource languages: Sorani Kurdish and Tajik. In *Proc. of the 26th International Conference on Computational Linguistics (COLING2016)*.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas W. Oard, and David Doermann. 2011. Cross-language entity linking. In *Proc. of 5th International Joint Conference on Natural Language Processing (IJCNLP2011)*.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 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)*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)*.

Avirup Sil and Radu Florian. 2016. One for all: Towards language independent named entity linking. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016)*.

Valentin I Spitkovsky and Angel X Chang. 2011. Strong baselines for cross-lingual entity linking. In *Proc. of the Text Analysis Conference (TAC2011)*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proc. of the 16th international conference on World Wide Web (WWW2017)*.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proc. of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL2016)*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *Proc. of the 5th International Conference on Learning Representations (ICLR2017)*.

Boliang Zhang, Xiaoman Pan, Tianlu Wang, Ashish Vaswani, Heng Ji, Kevin Knight, and Daniel Marcu. 2016. Name tagging for low-resource incident languages based on expectation-driven learning. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*.