Multi-lingual Entity Discovery and Linking

Avirup Sil
IBM Research AI
Yorktown Heights, NY
avi@us.ibm.com

Heng Ji
Rensselaer Polytechnic Institute
Troy, NY
jih@rpi.edu

Dan Roth
University of Pennsylvania
Philadelphia, PA
danroth@seas.upenn.edu

Silviu Cucerzan
Microsoft Research
Redmond, WA
silviu@microsoft.com

1 Overall

We live in a golden age of information, where we have access to vast amount of data in various forms: text, video and audio. Over the last few years, one of the key task that has been studied in support of natural language understanding and information extraction from text, is the task of Entity Linking (previously studied as Wikification). Entity Linking (henceforth, EL) (Bunescu and Pasca, 2006; Cucerzan, 2007; Ratinov et al., 2011) is the task of mapping mentions of entities in a text document to an entry in a large catalog of entities such as Wikipedia or another knowledge base (KB). It has also been one of the major tasks in the Knowledge-Base Population track at the Text Analysis Conference (TAC) (McNamee and Dang, 2009b; Ji and Grishman, 2011; Ji et al., 2014). Most works in the literature have used Wikipedia as this target catalog of entities because of its wide coverage and its frequent updates made by the community. The previous Entity Linking tutorial in ACL 2014 (Roth et al., 2014) addressed mostly EL research which have focused on English, the most prevalent language on the web and the one with the largest Wikipedia datasets. However, in the last few years research has shifted to address the EL task in other languages, some of which have very large web presence, such as Spanish (Fahrni et al., 2013; Ji et al., 2014), and Chinese (Cao et al., 2014; Shi et al., 2014) but also in others. In particular, there has been interest in cross-lingual EL (Tsai and Roth, 2016; Sil and Florian, 2016): given a mention in a foreign language document, map it to the corresponding page in the English Wikipedia. Beyond the motivation that drives the English EL task – knowledge acquisition and information extraction – in the cross-lingual case and especially when dealing with low resource languages, the hope is to provide improved natural language understanding capabilities for the many languages for which we have few linguistic resources and annotation and no machine translation technology. The LoreHLT2016-2017 evaluation\(^1\) and TAC 2017 pilot evaluation\(^2\) target really low-resource languages like Northern Sotho or Kikuyu which only have about 4000 Wikipedia pages (about 1/1000 the size of the English wikipedia).

The primary goals of this tutorial are to review the framework of cross-lingual EL and motivate it as a broad paradigm for the Information Extraction task. We will start by discussing the traditional EL techniques and metrics and address questions relevant to the adequacy of these to across domains and languages. We will then present more recent approaches such as Neural EL, discuss the basic building blocks of a state-of-the-art neural EL system and analyze some of the current results on English EL. We will then proceed to Cross-lingual EL and discuss methods that work across languages. In particular, we will discuss and compare multiple methods that make use of multi-lingual word embeddings. We will also present EL methods that work for both name tagging and linking in very low resource languages. Finally, we will discuss the uses of cross-lingual EL in a variety of applications like search engines and commercial product selling applications. Also, contrary to the 2014 EL tutorial, we will also focus on Entity Discovery which is an essential component of EL.

The tutorial will be useful for both senior and junior researchers (in academia and industry) with interests in cross-source information extraction and linking, knowledge acquisition, and the use of acquired knowledge in natural language processing and information extraction. We will try to provide a concise road-map of recent approaches, perspectives, and results, as well as point to some of our EL resources that are available to the research community.

\(^1\)https://lorehlt.nist.gov/

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts, pages 1–8
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics
2 Brief Tutorial Outline

2.1 Motivation and Overview [20 mins]

We will motivate the general EL problem (for English) by teaching the general methods that incorporate distance measures (Ratinov et al., 2011; Sil and Yates, 2013; Cheng and Roth, 2013). We will then briefly discuss multi-lingual IE problems and motivate cross-lingual EL (Ji et al., 2014; Sil and Florian, 2016). Then we will motivate the new trend of modeling distributional representations instead of distance.

2.2 Key Challenges and Multi-lingual Embeddings [20 mins]

We will present some key challenges daunting high-performing traditional EL systems and candidate generation and transliteration (Tsai and Roth, 2018) from a knowledge-base. We will also present the models for traditional cross-lingual EL (Sil and Florian, 2016; Tsai and Roth, 2016) and discuss some of their challenges: matching context between non-English documents with the English Wikipedia. Recently, neural Entity Discovery and Linking (henceforth, EDL) techniques have combated some of these challenges. These systems use multi-lingual embeddings which are essential building blocks for these neural architectures. Hence, before diving into the architectures we will survey multi-lingual embedding techniques (Mikolov et al., 2013c; Faruqui and Dyer, 2014; Ammar et al., 2016) and which ones work best for neural EL systems and motivate neural EL.

2.3 Neural Methods for EDL [30 mins]

Various shared tasks such as TAC-KBP, ACE and CONLL, along with corpora like OntoNotes and ERE have provided the community substantial amount of annotations for both entity mention extraction (1,500+ documents) and entity linking (5,000+ query entities). Therefore supervised models have become popular again for each step of EDL. Among all of the supervised learning frameworks for mention extraction, the most popular one is a combined Deep Neural Networks architecture consisted of Bi-directional Long Short-Term Memory networks (Bi-LSTM) (Graves et al., 2013) and CRFs (Lample et al., 2016). In TAC-KBP2017 many teams trained this framework from the same training data (KBP2015 and KBP2016 EDL corpora) and the same set of features (word and entity embeddings), but got very different results. The mention extraction F-score gap between the best system and the worst system is about 24%. We will provide a systematic comparison and analysis on reasons that cause this big gap. We will also introduce techniques to make the framework more robust to noise in low-resource settings.

We will then teach neural EL architectures (Globerson et al., 2016; Gupta et al., 2017a; Sil et al., 2018) that can tackle some of the challenges of the traditional systems. Then we will proceed to cross-lingual neural EL and survey the pipelines that most of these EL systems employ: cross-lingual NER and in-document coreference resolution. We will talk about how to model the contexts using various neural techniques like CNNs, LSTMs etc. and how systems compute similarity metrics of varying granularity (Francis-Landau et al., 2016; Sil et al., 2018).

2.4 Coffee break: [30 minutes]

2.5 Language Universal Methods for Cross-lingual EDL [30 mins]

We will then present some recent advances at developing low-cost approaches to perform cross-lingual EL for 282 Wikipedia languages, such as deriving silver-standard annotations by transferring annotations from English to other languages through cross-lingual links and KB properties, refining annotations through self-training and topic selection, deriving language-specific morphology features from anchor links, and mining word translation pairs from cross-lingual links (Pan et al., 2017a). We will also introduce some recent extensions along this line of work, including extending the number of entity types from five to thousands, and its impact on other NLP applications such as Machine Translation.

2.6 Multiple Knowledge Bases [25 mins]

A task that is similar to multi-lingual EL in both definition and approaches is domain-specific linking of entities in documents based on a given set of domains/corresponding knowledge repositories (Gao and Cucerzan, 2017). This task is important for applications such as the analysis and indexing of corporate document repositories, in which many of the entities of interest are not part of the general-knowledge but are rather company-specific and may need to be kept private. Conflating such terminologies and knowledge into one single knowledge model would be both daunting and undesirable. Thus, similarly to handling multiple languages, a system built for multiple-
domain linking, has to model each domain separately. We will discuss a multi-KB entity linking framework that employs one general-knowledge KB and a large set of domain-specific KBs as linking targets that extends the work from (Cucerzan, 2007, 2014a), as well as a supervised model with a large and diverse set of features to detect when a domain-specific KB matches a document targeted for entity analysis (Gao and Cucerzan, 2017).

2.7 New Tasks, Trends and Open Questions [15 mins]

Here, we will address some of the new settings: multi-lingual EL for search engines (Pappu et al., 2017; Tan et al., 2017). We will discuss some open questions such as improving the title candidate generation process for situations where the corresponding titles only exist in the English Wikipedia and also investigate the topological structure of related languages and exploit cross-lingual knowledge transfer to enhance the quality of extraction and linking (Tsai and Roth, 2018). We will also discuss EL for noisy data like social media (Meij et al., 2012; Guo et al., 2013). Finally, we will discuss the possibilities of extending the ideas taught in this EL tutorial to other multi-lingual IE tasks.

2.8 System Demos and Resources [10 mins]

Finally, we will show some demos of multi-lingual EL systems from the industry and academia. We will also provide pointers to resources, including data sets and software.

3 Tutorial Instructors

- **Name:** Avirup Sil  
  **Affiliation:** IBM Research AI  
  **Email:** avi@us.ibm.com  
  **Website:** Avi’s Webpage  
  Avi is a Research Staff Member and the chair of the NLP community at IBM Research AI. His research interests are in multi-lingual information extraction from large text collections (cross-lingual entity extraction, disambiguation and slot filling), machine learning and knowledge representation. Avi has published several papers on Entity Linking and his systems at IBM have obtained top scores in TAC-KBP annual multi-lingual entity linking evaluations. Avi is an area chair for Information Extraction at NAACL 2018 and also for COLING 2018. He is also organizing the workshop on the “Relevance of Linguistic Structure in Neural NLP” at ACL 2018.

- **Name:** Heng Ji  
  **Affiliation:** Rensselaer Polytechnic Institute  
  **Email:** jih@rpi.edu  
  **Website:** Heng’s Webpage  
  Heng Ji is the Edward G. Hamilton Development Chair Professor in Computer Science Department of Rensselaer Polytechnic Institute. Her research interests focus on Natural Language Processing, especially on Cross-source Information Extraction and Knowledge Base Population. She coordinated the NIST TAC Knowledge Base Population task since 2010 and has published many papers on entity discovery and linking. Heng has co-taught the “Wikification and Beyond: The Challenges of Entity and Concept Grounding” tutorial with Dan Roth at ACL 2014.

- **Name:** Dan Roth  
  **Affiliation:** University of Pennsylvania  
  **Email:** danroth@seas.upenn.edu  
  **Website:** Dan’s Webpage  
  Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, University of Pennsylvania. He is a fellow of AAAS, AAAI, ACL, and the ACM and the winner of the IJCAI-2017 John McCarthy Award, for “major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning.” Roth has published broadly in machine learning, natural language processing, knowledge representation and reasoning, and has developed several machine learning based natural language processing systems that are widely used in the computational linguistics community and in industry. Over the last few years he has worked on Entity Linking and Wikification. He has taught several tutorials at ACL/NAACL/ECL and other forums. Dan has co-taught the “Wikification and Beyond: The Challenges of Entity and Concept Grounding” tutorial with Heng Ji at ACL 2014.

- **Name:** Silviu Cucerzan  
  **Affiliation:** Microsoft Research  
  **Email:** silviu@microsoft.com  
  **Website:** Silviu’s Webpage  
  Silviu Cucerzan is a Principal Researcher at Microsoft Research and the Bing Knowledge Graph group. His research has focused on topics at the intersection of NLP
and IR with concrete applications to industry, including multilingual spelling correction, question answering, entity recognition and linking, query suggestion, vertical search, and ads selection. Many of the technologies developed by Silviu have been shipped with Microsoft products. The NEMO entity linking system developed by Silviu has scored the top performance during the four consecutive years it participated in the TAC-KBP evaluations organized by NIST and LDC.

References


Ziqiang Cao, Sujian Li, and Heng Ji. 2014. Joint learning of chinese words, terms and keywords. In EMNLP.


Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. ACL.


Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amanagn Subramanya, Michael Ringgaard, and Fernando Pereira. 2016. Collective entity resolution with multi-focal attention. ACL.


Stephen Guo, Ming-Wei Chang, and Emre Kıcıman. 2013. To link or not to link? a study on end-to-end tweet entity linking. In NAACL.


X. Han, L. Sun, and J. Zhao. 2011a. Collective entity linking in web text: a graph-based method. In SIGIR.


Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the tac2011 knowledge base population track. TAC.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In TAC.

Alex Judea, Benjamin Heisterling, Angela Fahrni, and Michael Strube. 2014. Hits monolingual and cross-lingual entity linking system at TAC 2014. In Text Analysis Conference.


the North American Chapter of the Association for Computational Linguistics Human Language Technologies.


Hailun Lin, Zeya Zhao, Yantao Jia, Yuanzhuo Wang, Jinhua Xiong, Xiaojing Li, Yuxiao Chang, Manling Li, Hongbo Xu, and Xueqi Cheng. 2014. Openkn at tac kbp 2014. In Text Analysis Conference.


Thomas Lin, Mausam, and Oren Etzioni. 2012b. Entity Linking at Web Scale. In AKBC-WEKEX.

Thomas Lin, Mausam, and Oren Etzioni. 2012c. No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. In EMNLP.


Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In WSDM.


Yuval Merhav, Joel Barry, James Clarke, David Murgatroyd, and One Alewife Center. 2013. Basis technology at tac 2013 entity linking.


Sean Monahan, Dean Carpenter, Maxim Gorelkin, Kevin Crosby, and Mary Brunson. 2014. Populating a knowledge base with entities and events. In Text Analysis Conference.

Thien Huu Nguyen, Yifan He, Maria Pershina, Xiang Li, and Ralph Grishman. 2014. New york university 2014 knowledge base population systems. In Text Analysis Conference.


Dan Roth, Heng Ji, Ming-Wei Chang, and Taylor Cascidy. 2014. Wikification and beyond: The challenges of entity and concept grounding. In ACL Tutorials.


Xing Shi, Kevin Knight, and Heng Ji. 2014. How to speak a language without knowing it. In ACL.


Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. 2012. Linking Named Entities to Any Database. In EMNLP-CoNLL.


Avirup Sil and Radu Florian. 2016. One for all: Towards language independent named entity linking. ACL.


Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013c. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP. volume 1631, page 1642.


Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016b. Sentence similarity learning by lexical decomposition and composition. *COLING*.


UK WR143PS. 1989. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters.
