Illinois Cross-Lingual Wikifier: Grounding Entities in Many Languages to the English Wikipedia

Chen-Tse Tsai and Dan Roth

Abstract

We release a cross-lingual wikification system for all languages in Wikipedia. Given a piece of text in any supported language, the system identifies names of people, locations, organizations, and grounds these names to the corresponding English Wikipedia entries. The system is based on two components: a cross-lingual named entity recognition (NER) model and a cross-lingual mention grounding model. The cross-lingual NER model is a language-independent model which can extract named entity mentions in the text of any language in Wikipedia. The extracted mentions are then grounded to the English Wikipedia using the cross-lingual mention grounding model. The only resources required to train the proposed system are the multilingual Wikipedia dump and existing training data for English NER.

Source code: github.com/cttsai/illinois-cross-lingual-wikifier
Demo: cogcomp.cs.illinois.edu/page/demo_view/xl_wikifier

Screen Shot

System Pipeline

- Tsai et al., CoNLL 2016. Cross-lingual Named Entity Recognition via Wikification
- This model can be trained on one or several languages, and can be applied to other languages directly
- The key idea is that the cross-lingual mention grounding model generates good language-independent NER features

- Tsai and Roth, NAACL 2016. Cross-lingual Wikification Using Multilingual Embeddings
- This model uses cross-lingual word and title embeddings to disambiguate the mentions extracted by the NER model to the English Wikipedia

Component Evaluation

- Cross-Lingual NER (F1 score)

	Dutch	German	Spanish	Turkish	Tagalog	Yoruba	Bengali	Tamil
Transfer from English	61.56	48.12	60.55	47.12	65.44	36.65	43.27	29.64
Monolingual	84.49	73.13	83.87	73.86	77.64	57.60	71.15	60.02

- Cross-Lingual Mention Grounding

	German	Spanish	French	Italian	Chinese	Hebrew	Thai	Arabic	Turkish	Tamil	Tagalog	Urdu
Prec@1	81.45	81.37	79.65	79.79	84.55	84.03	89.46	86.13	85.10	84.15	84.54	91.07

End-to-End System Evaluation

TAC 2016 Entity Discovery and Linking Tasks

Metric	Spanish	Chinese
Strong mention match	83.8	78.9
Strong typed mention match	81.3	75.9
Strong typed all match	73.3	68.2

This research is supported by NIH grant U54-GM114838, a grant from the Allen Institute for Artificial Intelligence (allenai.org), and Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.)