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

Chen-Tse Tsai and **Dan Roth**
University of Illinois at Urbana-Champaign
201 N. Goodwin, Urbana, Illinois, 61801
{ctsai12, danr}@illinois.edu

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. The system is online at http://cogcomp.cs.illinois.edu/page/demo_view/xl_wikifier

1 Motivation

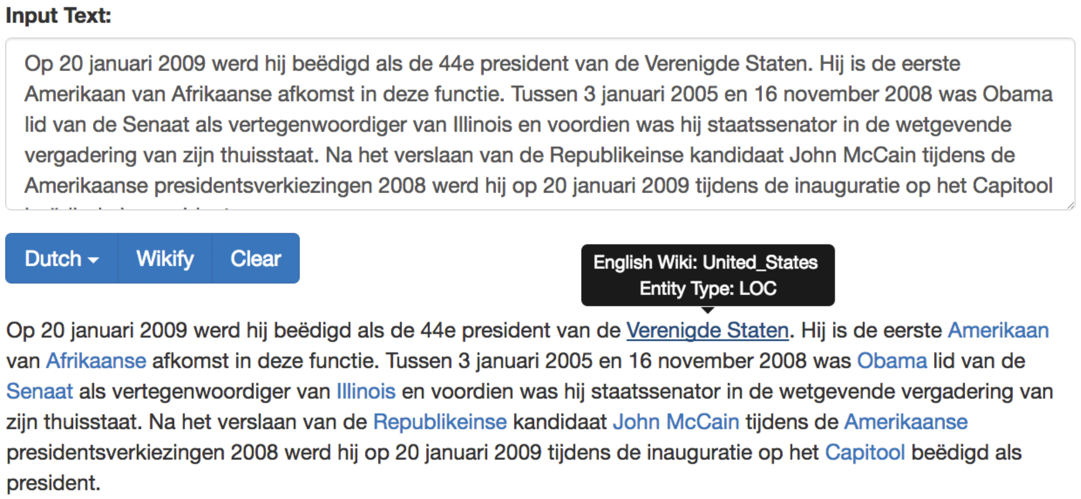
Wikipedia has become an indispensable resource in knowledge acquisition and text understanding for both human beings and computers. The task of wikification or Entity Linking aims at disambiguating mentions (sub-strings) in text to the corresponding titles (entries) in Wikipedia. For English text, this problem has been studied extensively. (Bunescu and Pasca, 2006; Cucerzan, 2007; Mihalcea and Csomai, 2007; Ratinov et al., 2011; Cheng and Roth, 2013) It also has been shown to be a valuable component of several natural language processing and information extraction tasks across different domains.

Recently, there has also been interest in the cross-lingual setting of Wikification: given a mention from a document written in a non-English language, the goal is to find the corresponding title in the English Wikipedia. This task is driven partly by the fact that a lot of information around the world may be written in a foreign language for which there are limited linguistic resources and, specifically, no English translation technology. Instead of translating the whole document to English, *grounding* the important entity mentions in the English Wikipedia may be a good solution that could better capture the key message of the text, especially if it can be reliably achieved with fewer resources than those needed to develop a translation system.

There are several language-specific Wikification systems but very few multilingual or cross-lingual systems. For instance, for English, there are Illinois Wikifier[1] and AIDA[2]. TagMe[3] supports Englsih, Italian, and German. RPI[4] developed systems for English, Spanish, and Chinese. However, even many widely spoken languages are not supported. Since these systems often use language-specific resources, it is hard to adapt these systems to a new language. The goal of the proposed system is to cover all 292 languages in Wikipedia and directly link mentions to the English Wikipedia titles using very little language-specific resource. That is, only using information in multilingual Wikipedia and some knowledge for tokenization, we are able to extract named entity mentions and ground them to the English Wikipedia for all languages in Wikipedia.

[1] http://cogcomp.cs.illinois.edu/page/demo_view/Wikifier
[2] http://github.com/yago-naga/aida
[3] http://tagme.d4science.org/tagme/
[4] http://blender04.cs.rpi.edu/~panx2/edl/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations,
pages 146–150, Osaka, Japan, December 11-17 2016.



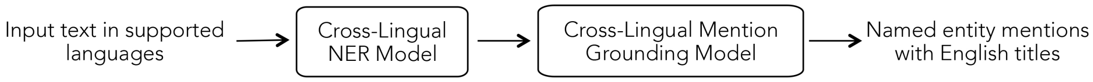
Input Text:

Op 20 januari 2009 werd hij beëdigd als de 44e president van de Verenigde Staten. Hij is de eerste Amerikaan van Afrikaanse afkomst in deze functie. Tussen 3 januari 2005 en 16 november 2008 was Obama lid van de Senaat als vertegenwoordiger van Illinois en voordien was hij staatssenator in de wetgevende vergadering van zijn thuisstaat. Na het verslaan van de Republikeinse kandidaat John McCain tijdens de Amerikaanse presidentsverkiezingen 2008 werd hij op 20 januari 2009 tijdens de inauguratie op het Capitool

English Wiki: United_States
Entity Type: LOC

Op 20 januari 2009 werd hij beëdigd als de 44e president van de Verenigde Staten. Hij is de eerste Amerikaan van Afrikaanse afkomst in deze functie. Tussen 3 januari 2005 en 16 november 2008 was Obama lid van de Senaat als vertegenwoordiger van Illinois en voordien was hij staatssenator in de wetgevende vergadering van zijn thuisstaat. Na het verslaan van de Republikeinse kandidaat John McCain tijdens de Amerikaanse presidentsverkiezingen 2008 werd hij op 20 januari 2009 tijdens de inauguratie op het Capitool beëdigd als president.

Figure 1: Screen shot of Illinois Cross-Lingual Wikifier.



Input text in supported languages
Cross-Lingual NER Model
Cross-Lingual Mention Grounding Model
Named entity mentions with English titles

Figure 2: Pipeline of Illinois Cross-Lingual Wikifier.

2 System Description

Figure 1 shows the web interface of our system. The bottom part is its output. The extracted named entities (in blue) are hyperlinked to the corresponding English Wikipedia pages. If the cursor points to a mention (e.g., "Verenigde Staten"), the corresponding English title and the entity type will be shown.

Our system is based on two components that we proposed recently: a cross-lingual NER model (Tsai et al., 2016) and a cross-lingual mention grounding model (Tsai and Roth, 2016). Figure 2 shows an overview of the system. Given some text in a non-English language, the cross-lingual NER model extracts named entity mentions and the cross-lingual mention grounding model finds the corresponding English Wikipedia titles for each mention.

2.1 Cross-Lingual Named Entity Recognition

We use the direct transfer NER model proposed in Tsai et al. (2016). This model can be trained on one or several languages, depending on the availability of training data, and can be applied to other Wikipedia languages without changing anything in the model. The key idea is that the cross-lingual mention grounding model (Section 2.2) generates good language-independent NER features for each word in any Wikipedia language. More specifically, by grounding all n-grams in the input text to the English Wikipedia, we can describe each word using a set of FreeBase types and Wikipedia categories. Since these FreeBase types and Wikipedia categories are always in English, the features extracted based on these types can be used across different languages.

The features used in our model include the standard lexical features, gazetteer features, and the features based on the cross-lingual mention grounding model. Note that as concluded in Tsai et al. (2016), we only use all features when the target language uses Latin script. Otherwise, only the language-independent features (based on the cross-lingual mention grounding model) are active. The model is trained on the English training data from CoNLL 2003 shared task. Therefore it follows the named entity definitions of the shared task which use four entity types: PER, ORG, LOC, and MISC.

Note that we use white spaces and few common punctuations to tokenize the input text for most languages. For the languages which need special tokenization or word segmentation, we try to find publicly available tokenizers. Otherwise, we simply treat each character as a token.

Approach	Dutch	German	Spanish	Turkish	Tagalog	Yoruba	Bengali	Tamil
Our system	**61.56**	**48.12**	**60.55**	**47.12**	**65.44**	**36.65**	**43.27**	**29.64**
Täckström et al. (2012)	58.4	40.4	59.3	-	-	-	-	-
Zhang et al. (2016)	-	-	-	43.6	51.3	36.0	34.8	26.0

Table 1: Performance of the cross-lingual NER model. The numbers are F1 scores. We compare our system with two related work which also assume no training data for the target language.

	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

Table 2: Evaluation of the cross-lingual mention grounding model on the Wikipedia dataset.

2.2 Cross-Lingual Mention Grounding

We adapt the model proposed in Tsai and Roth (2016), which uses cross-lingual word and title embeddings to disambiguate the mentions extracted by the NER model to the English Wikipedia. The model consists of two steps:

Candidate Generation: The first step is to select a set of English title candidates for each foreign mention. The goal of this step is to produce a manageable number of candidates so that a more sophisticated algorithm can be applied to disambiguate them. This step is achieved by dictionaries built from the hyperlink structure and inter-language links in Wikipedia. That is, for each English title, we gather all possible strings in any language that can be used to refer to it.

Note that the limitation of this procedure is that it only retrieves titles that are in the intersection of the English Wikipedia and the target language Wikipedia. That is, the English titles that are linked to some titles in the target language Wikipedia. For example, since *Dan Roth* does not have a page in the Chinese Wikipedia, this process will not generate *Dan Roth*'s English Wikipedia page as a candidate when we see his name in Chinese. To overcome this limitation, we extend the candidate generation process with a transliteration model (Pasternack and Roth, 2009). The model is trained on the (target language name, English name) pairs obtained from the Wikipedia titles. If the original candidate generator fails to retrieve any candidate for a mention, we transliterate the mention into English and then query title candidates by this English transliteration.

Candidate Ranking: Given a mention and a set of English title candidates, we compute a score for each title which indicates how relevant the title is to the mention. We represent a pair of (mention, candidate) by a set of features which are various similarities between them. These features are computed based on cross-lingual word and title embeddings. We embed words and Wikipedia titles of English and the target language into the same semantic space. By representing the mention using several contextual clues, we can compute meaningful similarity between the mention and a English title using the cross-lingual embeddings. We train a linear ranking SVM model to combine these features for each language. The training examples are constructed from the hyperlinked phrases in Wikipedia articles.

2.3 Evaluation

Since our goal is to have broad coverage of languages, we try to evaluate our system on as many languages as possible. However, only a couple of languages have end-to-end wikification datasets that also follow the CoNLL named entity definitions. Therefore, we evaluate the two key components separately.

The cross-lingual NER model is evaluated on 8 lanugages and the results are shown in Table 1. For Dutch, German, and Spanish, we use the test data from CoNLL 2002/2003 shared tasks. The data for the other five low-resource languages are from the LORELEI and REFLEX packages[5]. Comparing

[5] LDC2015E13, LDC2015E90, LDC2015E83, and LDC2015E91

Approach	Spanish	Chinese
Top TAC'15 systems	80.4	83.1
Our System	80.93	83.63

Table 3: Performance of the cross-lingual mention grounding model on TAC 2015 Entity Linking diagnostic task (mentions are given as the input). The numbers are precision@1.

our system to two related work which also assume no training data for the target language, our system outperforms them on all 8 languages.

The cross-lingual mention grounding model is evaluated on the Wikipedia dataset of 12 languages created by Tsai et al. (2016). Results are listed in Table 2. In this dataset, since at least one third of the query mentions cannot be solved by the most common title, the baseline that predicts the most common title has precision@1 at most 66.67 for each language. We can see that our system is much better than this baseline. Table 3 compares our system with the top systems participated in TAC 2015 Entity Linking diagnostic tasks. Our system achieves slightly better scores than the best systems of Spanish and Chinese.

3 Related Work

Besides the language-specific systems that we discussed in Section 1, Babelfy[6] has the most similar goals to our system. Babelfy grounds words and phrases of 271 languages to BabelNet, a multilingual encyclopedic constructed from multiple resources, including Wikipedia, WordNet, VerbNet, and so on. For each mention, it also provides the corresponding English entry if there is any linked to the grounded target language entry. Therefore Babelfy can be viewed as a cross-lingual grounding system.

The main differences between our system and Babelfy are two folds. First, the target mentions are different. In our system, we focus on grounding named entities thus there is a multilingual named entity recognition module. Babelfy tries to disambiguate all linkable n-grams. For example, the string "president of United States" is linked to Wikipedia titles president_of_United_States, United_States, president, and United_(Phoenix_album), and the word "state" is linked to a sense in WordNet. From this example, we can see that grounding words inside a name entity may not be very useful. Second, Babelfy only grounds mentions to the entries in the target language, and shows the corresponding English entry if the English entry is linked to the target language entry. As discussed in Section 2.2, these inter-language links could be very sparse for many languages. Instead, our system uses a transliteration model to retrieve English candidates from the foreign mentions directly. Moreover, due to cross-lingual word and title embeddings, our ranking model can directly compute similarities between foreign mentions and English titles.

4 Conclusion and Future Work

We release Illinois Cross-Lingual Wikifier, a tool that extracts named entities for many languages, and also grounds the extracted entities to the English Wikipedia. The broad coverage of our system will help people and computers to understand text in many languages especially when machine translation technology is unavailable or unreliable.

There are various directions which one can pursue to improve our system. While the cross-lingual NER model covers many languages, its performance still can be improved significantly. One possible direction is to select a better or closer source language for each target language. Another is to incorporate target language specific knowledge into the model. For the cross-lingual mention grounding model, the candidate generation does not have good enough coverage for small languages. Currently we simply use an off-the-shelf transliteration model to retrieve possible English titles. While this works well for people's names, many organization and location names need to be translated instead of transliterated.

[6] `http://babelfy.org/`

Acknowledgments

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.

References

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Chapter of the ACL (EACL).*

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP*, pages 708–716.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 233–242.

Jeff Pasternack and Dan Roth. 2009. Learning better transliterations. In *Proc. of the ACM Conference on Information and Knowledge Management (CIKM)*, 11.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL).*

Oscar Täckström, Ryan T. McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL.*

Chen-Tse Tsai and Dan Roth. 2016. Concept grounding to multiple knowledge bases via indirect supervision. *Transactions of the Association for Computational Linguistics*, 2.

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

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 *NAACL.*