Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Towards zero-shot cross-lingual named entity disambiguation

Ander Barrena [a], Aitor Soroa [a], Eneko Agirre [a]

[a] HiTZ Basque Center for Language Technologies – Ixa NLP Group, University of the Basque Country UPV/EHU, Donostia, Basque Country, Spain

ARTICLE INFO

Keywords:
Cross-lingual named entity disambiguation
Cross-lingual entity linking
Zero-shot learning
Transfer learning
Pre-trained language models
Low-resource languages

ABSTRACT

In cross-Lingual Named Entity Disambiguation (XNED) the task is to link Named Entity mentions in text in some native language to English entities in a knowledge graph. XNED systems usually require training data for each native language, limiting their application for low resource languages with small amounts of training data. Prior work have proposed so-called zero-shot transfer systems which are only trained in English training data, but required native prior probabilities of entities with respect to mentions, which had to be estimated from native training examples, limiting their practical interest. In this work we present a zero-shot XNED architecture where, instead of a single disambiguation model, we have a model for each possible mention string, thus eliminating the need for native prior probabilities. Our system improves over prior work in XNED datasets in Spanish and Chinese by 32 and 27 points, and matches the systems which do require native prior information. We experiment with different multilingual transfer strategies, showing that better results are obtained with a purpose-built multilingual pre-training method compared to state-of-the-art generic multilingual models such as XLM-R. We also discovered, surprisingly, that English is not necessarily the most effective zero-shot training language for XNED into English. For instance, Spanish is more effective when training a zero-shot XNED system that disambiguates Basque mentions with respect to an English knowledge graph.

1. Introduction

Information Extraction (IE) is the task of extracting structured information (company activities, medical records, etc.) from unstructured text. Early IE practitioners immediately noticed the importance of recognizing and identifying the named entities that appear in documents, such as person, organizations, locations, etc. However, named entity identification is a hard task that must overcome two main problems. On the one hand, entities can be referred to using many surface forms ("Barack Obama", "President Obama", "Mr. Obama", etc). On the other hand, entity mentions are often ambiguous. For instance, according to English Wikipedia the mention "Paul Newman" can refer to seven different entities, including the famous actor, but also a linguist, or even a rock band with the same name.[1]

The task that addresses the aforementioned problems is called Named Entity Disambiguation (NED), and its goal is to ground entity mentions in documents with entries of a knowledge-base (KB). NED is a fundamental task of semantic Web annotation with many downstream applications such as text mining (Derczynski et al., 2015) or authorship disambiguation (Veloso et al., 2012), to name a few. Most of the work in NED has been monolingual, where both the documents and the Knowledge Base are on the same language. In cross-lingual NED (XNED), however, documents are written in any language, and the mentions are linked to a foreign-language Wikipedia, typically English (McNamee, Mayfield, Lawrie, Oard, & Doermann, 2011; Tsai & Roth, 2016; Sil, Kundu, Florian, & Hamza, 2018; Zhao, Wu, Wang, & Li, 2016; Upadhyay, Gupta, & Roth, 2018). Cross-lingual NED has gained attention in the past years, as it allows extracting structured information from foreign languages with limited resources —or, even, no resources at all— and no machine translation technology. Linking the entities mentioned on documents to English Wikipedia is an important step towards effective informaton extraction in such languages. Fig. 1 shows an example in Basque, where a monolingual NED system links the target mention to the corresponding entity in the Basque Wikipedia, and a cross-lingual NED system links the mention to the English Wikipedia. In this case, the Basque mention *AEB*, which is the acronym for *Ameriketako Estatu Batuak* (*United States of America*) corresponds to the English target entity *United_States*.

NED is usually accomplished using supervised systems that require a high amount of annotated data containing documents where the entity mentions are manually linked to KB entries. Wikipedia is the natural choice for training, as editors have manually added hyperlinks to

E-mail addresses: ander.barrena@ehu.eus (A. Barrena), a.soroa@ehu.eus (A. Soroa), e.agirre@ehu.eus (E. Agirre).

[1] See https://en.wikipedia.org/wiki/Paul_Newman_(disambiguation)

https://doi.org/10.1016/j.eswa.2021.115542
Received 22 February 2021; Received in revised form 29 June 2021; Accepted 30 June 2021
Available online 9 July 2021

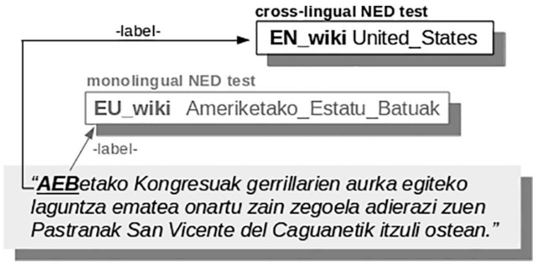
cross-lingual NED test
-label-
EN_wiki United_States
monolingual NED test
EU_wiki Ameriketako_Estatu_Batuak
-label-
"AEBetako Kongresuak gerrillarien aurka egiteko laguntza ematea onartu zain zegoela adierazi zuen Pastranak San Vicente del Caguanetik itzuli ostean."

Fig. 1. Example of monolingual and cross-lingual NED in Basque, where the target entity mention in Basque, ***AEB***, has to be disambiguated to the correct Wikipedia entities ***Ameriketako_Estatu_Batuak*** and ***United_States***, respectively.

articles, where the anchor text corresponds to the mention, and the url corresponds to the entity. Note that Wikipedia entries are routinely used in knowldge graphs such as Wikidata, DBpedia or BabelNet, among others (Pellissier Tanon, Vrandečić, Schaffert, Steiner, & Pintscher, 2016; Bizer et al., 2009; Navigli & Ponzetto, 2012). XNED systems are trained similarly to their monolingual counterparts, with the difference that the output entity is from a KB in another language (Tsai & Roth, 2016; Sil et al., 2018; Upadhyay et al., 2018; Rijhwani, Xie, Neubig, & Carbonell, 2019). The English Wikipedia contains millions of training examples, but, unfortunately, Wikipedias in many languages are much smaller. This lack of training data severely hinders the development of NED systems for low resource languages, both monolingual and cross-lingual.

As a solution to the small amounts of data in some languages, *transfer learning* techniques allow leveraging training data from one language to enhance the performance of a model on a different language (Ruder, 2019). Cross-lingual transfer learning (Smith, Turban, Hamblin, & Hammerla, 2017, 2018, 2020, 2019) allows to develop XNED models that are trained with data from resource-rich languages, and applied on a low resource language. In fact, several systems (Sil et al., 2018; Upadhyay et al., 2018) use English training data to train XNED models for other languages, without the need of native training examples, and claim to perform *zero-shot* transfer learning.

Although these systems (Sil et al., 2018; Upadhyay et al., 2018) do not directly access native training examples, they both need to combine the output of their systems with native prior probabilities for good performance. Given a mention string, these priors probabilities capture the probability distribution of entities for that mention, regardless of the context it appears. In the example above, the mention "Paul Newman", when used in Wikipedia, refers most of the time to the famous actor, and less often the politician, the linguist, or the rock band with the same name. Prior probabilities need to be calculated from training data, by just counting how many times each mention-entity pair occurs. Zero-shot approaches avoid the use of native training, and thus a system using entity priors cannot be considered to be zero-shot. However, both Sil et al. (2018) and Upadhyay et al. (2018) use such priors, and, for instance, Upadhyay et al. (2018) reports more than 25 points of accuracy drop when native priors are not used.

In this paper we present a zero-shot cross-lingual system that obtains good results without using native priors. Our system follows the so-called *word expert* approach presented in Barrena, Soroa, and Agirre (2018), which we adapt to the cross-lingual scenario. The system breaks the NED task into many classification tasks, one for each target mention string, e.g. "Paul Newman" or "AEB" (in Basque). Our XNED system builds a classifier for each entity mention in the native language which returns the intended entity in English Wikipedia. The system can be trained on English examples alone, and disambiguate mentions in other languages in a zero-shot fashion. We tried different transfer strategies and show the best results for a NED-oriented multilingual pre-training strategy, which obtains better results than the state-of-the-art in multilingual masked language models (XLM-R (Conneau & Lample, 2019)).

We performed XNED experiments in two high-resource languages (Spanish and Chinese) and one low-resource language (Basque). Given the scarcity of datasets for XNED, we present a new dataset with news documents in Basque manually linked to English Wikipedia.

The main contributions of our work are:

- A XNED system that does not need native priors, and which significantly improves state-of-the art results in resource-rich languages like Spanish and Chinese, and can be effectively applied to low-resource languages like Basque.
- Experiments testing different multilingual transfer strategies, showing that better results are obtained with a purpose-built multilingual pre-training method compared to state-of-the-art generic multilingual models such as XLM-R.
- Experiments that show that English is not necessarily the most effective training language for zero-shot XNED. For instance, training XNED models in Spanish results in better performance Basque XNED, even if the returned entities are from the English Wikipedia.
- A new dataset in Basque for XNED into English.

The rest of the paper is structured as follows. We first present related work, followed by our cross-lingual NED system. Section 4 presents the experimental settings and resources. Section 5 reviews the development experiments, which were carried out following the native setting of cross-lingual NED. Section 6 presents the results of the main experiments. Section 7 presents the comparison to the state of the art. Finally, the conclusions and future work. The cross-lingual word expert model and the Basque cross-lingual NED dataset are publicly available for reproducibility.[2]

2. Related work

The first cross-lingual NED systems were developed within the TAC-KB Entity Linking challenge starting in 2011 (Ji, Grishman, & Dang, 2011; Ji, Nothman, Hachey, & Florian, 2015) where participants had to link Spanish and Chinese documents to entities of English Wikipedia. Early cross-lingual systems either found the entities in the native language and then translated them to the target language, or relied on automatically translated queries to English and then performed English monolingual NED.

In zero-shot XNED, systems trained with examples in one language are applied to another language directly (Upadhyay et al., 2018; Sil et al., 2018; Rijhwani et al., 2019). These systems rely on cross-lingual embeddings, which represent words of different languages in the same shared space. Cross-lingual embeddings are usually built by independently training word embeddings in different languages, and mapping them to a shared space through linear transformations (Mikolov, Le, & Sutskever, 2013; Artetxe, Labaka, & Agirre, 2018). While early systems required bilingual dictionaries to learn the mapping, further work on unsupervised cross-lingual embeddings eliminated this requirement (Artetxe, Labaka, & Agirre, 2018). More recently, multilingual contextualized word embeddings such as multilingual BERT or XLM-R (Conneau et al., 2020) have proven to outperform static embeddings in many cross-lingual tasks. These models are pre-trained using corpora composed of documents in many languages, and often they do not require the documents to be aligned.

In one of the first uses of cross-lingual embeddings, Tsai and Roth (2016) present a XNED and cross-lingual Wikification system that uses a set of cross-lingual representations based on context words, entity titles and mention strings. It computes similarity scores among those representations to train a linear SVM algorithm. For each language, it trains a

[2] https://github.com/anderbarrena/xNED

monolingual entity and word representation model based on its corresponding Wikipedia, replacing anchor occurrences by its corresponding entity. Further, builds a multilingual entity and word representation model joining monolingual representations using Wikipedia interlanguage links as a seed dictionary. The system requires native examples, though, and does not perform zero-shot XNED.

Upadhyay et al. (2018) report the first zero-shot XNED results. They present a single neural model that encodes sentence in various languages using a CNN based encoder and cross-lingual word embeddings. The model is trained using a loss function that also incorporates the probability of a mention referring to an entity, as well as the types of the candidate entities. The final score is combined with native prior probabilities, and the performance of the system suffers significantly when these priors are not available, with two digit drops in performance. In contrast, instead of a building a single model for all mentions in all languages, our system trains many small models (one model per mention in each language), which yields better results overall. Moreover, we show that our approach is much more robust in the absence of native priors.

Sil et al. (2018) introduce a deep neural cross-lingual entity linking system using an ensemble composed of a variety of complex neural architecture combinations. The model combines both local and global algorithms using in-domain news data to train the model. Their zero-shot learning approach uses both native priors and supervised embedding mappings, and unfortunately they do not report figures without the native supervision. We show that our system obtains better results overall, and that in some languages (e.g. Spanish) our zero-shot system without entity priors performs better than theirs.

Table 1 presents the main characteristics of state-of-the-art XNED systems, including ours, and shows the main differences among them, including whether they are zero-shot, and whether they can perform zero-shot without using native priors. In addition, most XNED systems use bilingual dictionaries to build the cross-lingual embeddings. In contrast, our method uses an unsupervised approach to learn the embeddings, which requires no bilingual dictionary (Artetxe et al., 2018). Using bilingual dictionaries usually improves the embedding quality, but we wanted to use as few bilingual information as possible, hence proposing a method that is easily ported to any language pair. Likewise, these XNED methods include a step where the model is fine-tuned with in-domain training data, whereas our method exclusively uses Wikipedia to train the models.

Zero-shot XNED systems require cross-lingual alias tables that links mentions in the target language to candidates entities, which are often derived from manually annotated data in Wikipedia and from interlanguage links between entities in different languages.[3] Logeswaran et al. (2019) drop this requirement and present a zero-shot system that relies exclusively on entity descriptions to generate the candidates. Unfortunately, the system is only for English. However, we think that their approach is complementary to ours and sets a future direction towards dropping the dependency of alias tables in our system. Rijhwani et al. (2019) present a zero-shot cross-lingual candidate generation system focused on low resource languages with no bilingual resource available. They first train a model on a closely-related high resource language using bilingual links to English, and transfer the model to the low resource language. Their method can be seen as a candidate generation rather than a full XNED system, as they do not disambiguate mention in context, but rather return the same entity regardless of context. The authors do not test their system on the standard cross-lingual NED datasets, and therefore we can not perform a valid comparison among the systems. In any case, obtaining entity mappings across languages and cross-lingual candidate generation is an interesting research line for low-resource languages (Zhou, Rijhwani, Wieting, Carbonell, & Neubig, 2020; Zhou, Rijhwani, & Neubig, 2019) complementary to ours.

Table 1
Characteristics of state-of-the-art XNED systems in the columns: *zero-shot* or not, able to do zero-shot without native priors (*w/o priors*), use of unsupervised cross-lingual word embeddings (*Unsupervised mapping*), *Wikipedia only* for systems trained exclusively in Wikipedia.

	zero-shot	w/o priors	Unsupervised mapping	Wikipedia only
Tsai and Roth (2016)				✓
Sil et al. (2018)	✓			
Upadhyay et al. (2018)	✓	✓		
xWE (ours)	✓	✓	✓	✓

Close related task to NED such as Entity Resolution (ER, also known as Entity deduplication) is the task of identifying different representations of the same real-world entities across databases. Kasai, Qian, Gurajada, Li, and Popa (2019) target low-resource setting to ER task, designing a transfer learning approach from a high-resource setting. This work shows that similar ideas to ours are also successful beyond XNED tasks.

3. Cross-lingual NED system

Our XNED system follows the well known “word expert” model in Word Sense Disambiguation (Agirre and Edmonds, 2007), where we build one classifier for each entity mention string. The model follows the architecture presented in Barrena et al. (2018) and adapt it to the cross-lingual setting. We also propose several changes to the original model, which improve its performance as shown in the experimental section. In the following subsections we describe the system of Barrena et al. (2018) along with the proposed improvements and its adaptation to the cross lingual task.

3.1. Word expert models

The system builds a classifier for each ambiguous mention m that occurs in a context c. For a given mention m, the corresponding classifier will compute $P_m(e|c)$, the probability of m linking to an entity e given the context c. Given a suitable representation of c, the classifier consists of two fully connected layers following a softmax layer whose output are the possible entities the mention m can be linked to.[4]

Two word expert models are built for each mention, each model being trained on different data. The first model, $P_m(e|c)_{\text{orig}}$ is trained on examples of the mention m linking to possible candidates. The second model $P_m(e|c)_{\text{aug}}$ is trained by also considering occurrences of mentions different from m that link to a candidate entity of m (for example, when building an expert for the mention *New York* the augmented model will also consider examples of the mention *NY*, as both mentions link to the entity `New_York`). The augmented model is specially useful with mentions that contain few examples.

The final score of the word expert is the multiplication of both models:

$$e = \underset{e}{\operatorname{argmax}} P_m(e|c)_{\text{orig}} * P_m(e|c)_{\text{aug}} \tag{1}$$

As stated before, we propose a set of improvements to the model presented here. First, instead of having hidden layers of fixed dimension (256 units in the original model), we follow the pyramidal rule (Masters,

[3] https://en.wikipedia.org/wiki/Help:Interlanguage_links

[4] Section 4.1 describes how to compute the candidate entities for a given mention.

1993) to dynamically set the dimension of the ith hidden layer d_i as a function of the sizes of the input ($\mathbf{x}_i$) and output vectors ($\mathbf{y}_i$):

$$d_i = \sqrt{\pi \cdot \text{len}(\mathbf{x}_i) \cdot \text{len}(\mathbf{y}_i)} \quad (2)$$

Preliminary experiments showed this pyramidal architecture performed considerably better on mentions with small training data (less than 100 training instances[5]). We also use the ELU activation function instead of the original ReLU.

Fine-tuning the parameters for each of the 500 k expert models would require a very large amount of time. In view of this, the authors in Barrena et al. (2018) propose to build several models that are trained using the same learning rate, and average the output of the models to obtain the final score.[6] We adopt the same strategy but train four models per mention instead of three, using different learning rates.[7] Preliminary experiments suggested that higher learning rates are useful for mentions with small training data and vice versa.

We used the 4% of the training data available for each word expert as development. For each of the proposed 4 learning rate schedules, we choose adam for optimization and we early stop when development accuracy drops for one epoch. Then we divide the initial learning rate by 10 and we continue training until accuracy drops for another epoch, and we finally choose the best performing model (gathering 4 models using original data and 4 models using augmented data for each mention). We regularize the word expert using dropout after context representation and each hidden layer, with a dropout probability set to 0.16. At test time, following previous work, we also average the results of each original and augmented models and we multiply them to obtain the final score. Regarding to training and tuning details, remarkably, we only used Wikipedia, we did not use any in domain data to tune our Word Expert models.

3.2. Representing the context

In Barrena et al. (2018) the authors try different alternatives to represent mention contexts, which are the input of the word expert classifiers described above. In the following, we consider $c = \{w_1, \ldots, w_N\}$ an input context of N words, using previous and next sentences when the sentence where the target mention occurs is too short (we set context length N to 64). In this paper we consider the following methods to represent mention contexts:

Bag of Words. The bag of words representation is obtained by averaging the word embeddings of the words that comprise the context, after replacing the target mention with a vector whose elements are all one. Word embeddings are trained from Wikipedia.

Pretrained NED model. We pretrain a single neural model for monolingual NED, which we use to encode input mentions in the word expert. Note that we do not use the predictions of the model, it is just used for encoding. The model is composed by a context encoder based on LSTMs followed by a linear layer and softmax classifier, and is jointly trained to predict most of the correct entities and mentions in Wikipedia. Due to memory constraints we limit the number of entities to the 256k most frequent ones. After the pretraining stage, the softmax layer is discarded and the output of the linear context layer is used to produce new context representations. The contexts are represented by the hidden state of the last time-step of the LSTM. In this paper we use a slight variation of the pretrained NED model (dubbed *Stacked LSTM*). We use two LSTM layers of 2048 units instead of one, adding a dropout layer between stacked LSTMs (dropout probability 0.16) and the context layer dimensionality is 512. We used pretrained word embeddings as input (see Section 4.1) and we also represent the mention words as a constant word vector of ones. When training we set aside 4% of the Wikipedia data for development. We trained the model using adam optimizer and learning rate set to $1e-4$ and we early stop when validation accuracy drops for one epoch.

Pretrained language model. Different from Barrena et al. (2018), we also use a pretrained masked neural language model to represent the context (Pretrained LM for now on). We use the XLM-R state-of-the-art multilingual neural language model (Conneau et al., 2020) that allows us to represent the mention contexts in many languages in the same space. Following usual practice, we use the embedding of the first dummy token to represent the whole context. Note that we do not update the XLM-R weights when training word experts, as it would lead to different set of weights for each word expert.

3.3. Constructing a XNED system: native and zero-shot settings

In XNED, mentions occurring in documents written in a certain language (henceforth, the native language) are linked to entries of a KB from a different language (the foreign language, in our case, English). In principle, any monolingual NED system can be easily ported to the cross-lingual scenario, simply finding the entity in the foreign KB which corresponds to the native entity using Wikipedia interlanguage links (also called lang-links for short), and thus building a cross-lingual mention-entity alias table. We call this approach the **native** approach, as in this case both training and test data are in the same language, although the output entity is in another language. This is the most common setting when training data in the native language is available. Fig. 2 shows some training examples for monolingual NED, as well as the native XNED approach.

However, in many low resources languages the training data is scarce or even non-existent. In these cases, the **zero-shot** approach is a valid alternative. In the zero-shot approach, the model is trained using documents in a foreign language (usually the language of the target entities in the KB), and tested on documents in the native language. The system described in the previous section can be easily ported to the zero-shot setting, provided in can represent the context of occurrence of the mentions in the test language in the same multilingual space as the training examples in another language. For instance, Ruder (2019) and Conneau et al. (2020) have shown that cross-lingual embeddings can represent words of two languages in the same shared semantic space effectively, with good results in several tasks. Cross-lingual embeddings can be static or contextual, and in this work we try both approaches for XNED, as follows.

Regarding **static embeddings**, we build cross-lingual embeddings using VecMap, an unsupervised offline method that requires no parallel data or bilingual dictionaries (Artetxe et al., 2018). Those bilingual embeddings are used as input for the bag of words and pretrained NED models described in the previous section. Note that, in the latter case, a different pretrained NED model is necessary for each language pair. Regarding **contextual embeddings**, we use XLM-R, a multilingual pretrained language model system described in the previous Section, and is thus able to represent contexts in all the languages considered in our experiments without any change in the architecture.

To obtain the training data for zero-shot learning, the most usual case is to focus on the foreign language (English in our case), but in some cases, it might be interesting to gather examples in a third language, which we will call pivot language. We thus have the native language (the language of the test mentions and contexts), the foreign language (that of the KB with the target entities), and the pivot language (the language of the training examples). We can assume that the pivot language is the foreign language (typically English), but in Section 6.3 we will see that other options might be effective.

In order to collect the training data from the Wikipedia in the pivot language, we first consider **mention sharing**, i.e., we gather pivot training examples of mentions that are spelled the same way in the native and pivot languages (e.g. “Paul Newman”). Note that mention sharing can be very limited, depending on the pair of languages. For

[5] We do not train models having less than 10 training instances

[6] They trained 3 models per mention with a constant learning rate of $1e^{-3}$

[7] $7e^{-3}, 5e^{-3}, 3e^{-3}, 1e^{-3}$

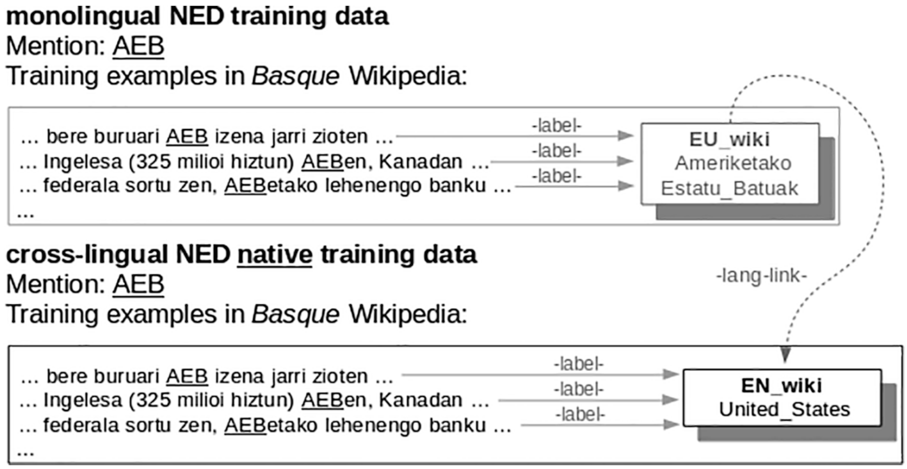
monolingual NED training data
Mention: AEB
Training examples in *Basque* Wikipedia:
... bere buruari AEB izena jarri zioten ...
... Ingelesa (325 milioi hiztun) AEBen, Kanadan ...
... federala sortu zen, AEBetako lehenengo banku ...
...
-label-
-label-
-label-
EU_wiki
Ameriketako
Estatu_Batuak
-lang-link-
cross-lingual NED native training data
Mention: AEB
Training examples in *Basque* Wikipedia:
... bere buruari AEB izena jarri zioten ...
... Ingelesa (325 milioi hiztun) AEBen, Kanadan ...
... federala sortu zen, AEBetako lehenengo banku ...
...
-label-
-label-
-label-
EN_wiki
United_States

Fig. 2. **Monolingual** NED training examples for the Basque mention string *AEB* (top). In the bottom, training examples for the **native** cross-lingual approach for the same mention string, derived from the monolingual examples using Wikipedia the lang-link (interlanguage link) between the Basque entity *Ameriketako_Estatu_Batuak* and the English couterpart *United_States*.

instance, for the mention “AEB” in Basque, we would only be able to collect training data for its “Advanced_Braking_System” meaning, see Fig. 3.

Alternatively, the **augmented model** presented in Section 3.1 can be used to gather examples of mentions in the pivot language that link to the same (pivot) entity of the original native mention. See Fig. 3 for an example. As a result of both mention sharing and the augmented model, our zero-shot system is able to collect training examples to train most of the mentions in the native language.

4. Experimental settings and resources

The xWE cross-lingual NED system has been evaluated in two resource-rich languages (Spanish and Chinese) and in a low-resource language (Basque). The input documents are written in those languages, and the entity mentions must be linked to English Wikipedia articles. We evaluate Spanish and Chinese on the TAC-KBP 2015 dataset (Ji et al., 2015), which comprises news documents from the Gigaword Corpus. Note that other evaluation datasets for XNED (Tsai & Roth, 2016; Upadhyay et al., 2018) are derived from Wikipedia, and thus, the training and test examples come from the same distribution. In contrast, the TAC-KBP datasets are from the News domain, and offer a more challenging and realistic evaluation dataset.

Regarding Basque evaluation, we built a new dataset called EusE, which is derived from the monolingual dataset presented in Fernandez (2012). EusE contains 1032 named entity mentions occurring on Basque News documents[8] that are manually linked to English Wikipedia entities.

Accuracy is the evaluation metric in all our experiments, the fraction of correctly disambiguated mentions divided by the total number of mentions that need to be linked to the KB. This measure is referred as *inKB* accuracy in TAC-KBP (Ji et al., 2015).

4.1. Resources

Our XNED model is trained using the 2014 Wikipedia dumps for English (en), Spanish (es), Chinese (zh) and Basque (eu). For each entity mention, we first build a so called cross-lingual *alias table* with the list of possible entities (Tsai & Roth, 2016; Upadhyay et al., 2018; Sil et al., 2018). Cross-lingual alias tables are built for each language by first building monolingual alias tables as in Barrena et al. (2018), and then using Wikipedia interlanguage links to obtain the corresponding English entities. For Chinese, we generated additional mentions using the heuristics proposed in Tsai and Roth (2016), breaking each Chinese mention string into tokens, and creating a new alias dictionary entry which maps tokens to English entities. We kept the 32 most frequent entities for each mention in all the alias tables, as in Barrena et al. (2018). This first step defines an upperbound for the XNED systems, and we present an analysis of the coverage of alias tables in Section 6.4.

Training data for each target mention string is obtained from Wikipedia examples, gathering all entity occurrences along with their contexts for the languages listed above, following usual practice (Tsai & Roth, 2016; Upadhyay et al., 2018; Sil et al., 2018; Barrena et al., 2018). Wikipedia dumps are preprocessed using a simple tokenizer except for Chinese Wikipedia, which was tokenized using the Stanford Word Segmenter (Chang, Galley, & Manning, 2008). The Wikipedia examples are split into training and development sets, and models that performs best on their respective development set are selected for testing. No additional training data is used to fine-tune our models.

Monolingual embeddings were trained using *fastText* (Bojanowski, Grave, Joulin, & Mikolov, 2017) in their respective Wikipedias. The embedding dimension is 512.[9] Cross-lingual embeddings were built using the unsupervised algorithm of VecMap (Artetxe et al., 2018) that requires no seed dictionary or parallel data to compute the cross-lingual mappings, making it easier to build cross-lingual embeddings for all language pairs. We kept all the embeddings frozen in all the experiments bellow. Regarding contextual embeddings, we use the base-size pretrained language model of XLM-R[10] as provided by the authors, which is trained on 2.5 TB of CommonCrawl data in 100 languages.

5. Development experiments on the native setting

In this section we describe the development experiments with the aim to evaluate the main variants of our NED model, as described in Section 3. In order to speed up development, we chose to perform the development experiments on the language with the smallest Wikipedia, Basque, and only explored the native approach, avoiding zero-shot experiments. In order not to touch the test data, the development was done entirely on the Basque Wikipedia. For each mention in the Basque EusE dataset, we collect all Wikipedia occurrences and use %96 of mentions for training and %4 for validation. Results are reported on the validation

[8] The documents are from the Basque newspaper “Euskaldunon Egunkaria”.

[9] We did not check other embeddings sizes.

[10] https://github.com/pytorch/fairseq/tree/master/examples/xlmr

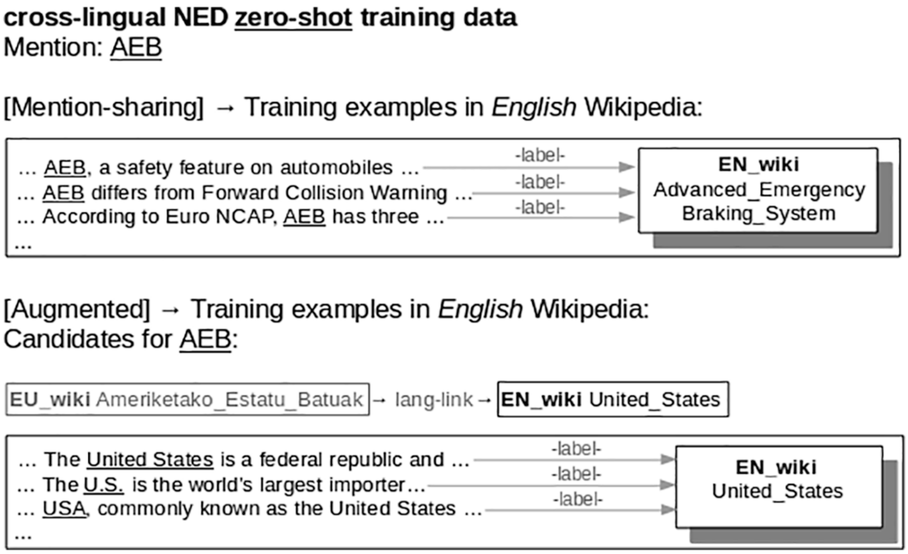
cross-lingual NED zero-shot training data
Mention: AEB
[Mention-sharing] → Training examples in English Wikipedia:
... AEB, a safety feature on automobiles ...
... AEB differs from Forward Collision Warning ...
... According to Euro NCAP, AEB has three ...
...
-label-
-label-
-label-
EN_wiki
Advanced_Emergency
Braking_System
[Augmented] → Training examples in English Wikipedia:
Candidates for AEB:
EU_wiki Ameriketako_Estatu_Batuak
→ lang-link →
EN_wiki United_States
... The United States is a federal republic and ...
... The U.S. is the world's largest importer...
... USA, commonly known as the United States ...
...
-label-
-label-
-label-
EN_wiki
United_States

Fig. 3. Zero-shot XNED training examples for the Basque mention string *AEB*, gathered using two methods. In **mention sharing** (top), English examples for the mention string *AEB* are gathered with their labels. In the **augmented model** (bottom), the Basque alias table for *AEB* lists Basque candidate entity *Ameriketako_Estatu_Batuak*, which has a lang-link (interlanguage link) to the English entity *United_States*. Therefore, English examples labeled with *United_States* are added to the pool of training examples for Basque mention *AEB*.

set for the *orig* model (see Section 3.1 for model details). For the sake of space, we omit the results for the *aug* model, as preliminary results showed that they follow the same trends.

Table 2 shows the results of incrementally incorporating the proposed modifications in the original system to all three representation models. All the modifications cause the results to improve in all context representation options (bag of words, pretrained NED model or pretraind XLM-R). The best results are obtained when incorporating all the modifications,[11] which shows that the partial gains of each change are complementary. Compared to the model of Barrena et al. (2018) in the first row, our cross-lingual Word Expert xWE (last row) improves the results by more that 5 points regardless of the method used for representing the context. Regarding the context representation options, the best results are for the pretrained NED contextual model, followed by XLM-R and bag-of-words, respectively. This confirms the results obtained in our previous approach in (Barrena et al., 2018), showing that pretrained NED also outperforms bag of words model in cross-lingual scenario. As we improved the pretrained NED design, here the improvement is even larger.

6. Results

In this section we report our main results for Spanish, Basque and Chinese, also comparing them with current state-of-the-art system in cross-lingual NED.

Table 2
Development experiments of native NED for Basque as accuracy. First row corresponds to the original model (Barrena et al., 2018), and successive rows show the results when including a single modification to the original model. The last row includes all three modifications. Best results in bold.

		Bag of Words	Pretrained NED	Pretrained XLM-R
native	Barrena et al. (2018)	91.12	93.24	91.45
	→ *pyramidal*	93.88	95.76	93.09
	→ *mixed lr*	92.77	97.01	94.76
	→ *stacked LSTM*	—	96.43	—
	→ *all* (xWE)	**96.54**	**98.42**	**97.07**

6.1. Native and zero-shot settings

Table 3 shows the main XNED results. The top section of the table correspond to the zero-shot setting and the bottom section to the native setting (c.f. Section 3.3). In these experiments English is used as foreign language, and also as pivot language in the zero-shot setting. The fist row on each part shows the results of a baseline system that assigns the most probable entity to each mention regardless of context, according to entity prior probabilities, which are calculated using native data in each language. The Table shows that this is a strong baseline for all datasets and especially for Basque. Note that we do not report prior baseline results for the zero-shot setting, as they are unavailable in that setting.

Regarding the context representation methods, they follow the trend observed in the previous Section, with NED pretraining being the best method. It consistently beats the bag-of-words representation in all datasets. It also beats XLM-R in most cases, with a difference of up to 3.4 points in Chinese (native setting), with the exception of Basque (in the zero-shot setting) with a small difference. The results show that the simple bag-of-words method for representing the context is a strong contender, outperforming XLM-R in both native and zero-shot. They also indicate that for this particular problem, building static embeddings from Wikipedia itself, and mapping them to a cross-lingual space is preferable than using multilingual pretrained language models, and are complementary to the literature reporting the contrary in other downstream tasks (Conneau et al., 2020). Following this result, we use the NED pretraining method for representing the context in the coming experiments.

Regarding zero-shot, the performance of zero-shot systems are below native systems, but they are remarkably close in Spanish and Chinese. In

Table 3
Zero-shot and native cross-lingual NED results as accuracy. Bold marks the best results in each scenario (native or zero-shot) per dataset and average.

		Basque	Spanish	Chinese	avg.
zero-shot	*Prior baseline*	—	—	—	—
	xWE (*bag of words*)	81.67	84.79	83.31	83.26
	xWE (*pretrained NED*)	82.17	**85.48**	**83.37**	**83.67**
	xWE (*pretrained XLM-R*)	**82.55**	83.14	83.08	82.92
native	*Prior baseline*	87.10	80.51	80.89	82.83
	xWE (*bag of words*)	91.66	85.91	86.13	87.90
	xWE (*pretrained NED*)	**92.41**	**87.60**	**87.09**	**89.03**
	xWE (*pretrained XLM-R*)	91.15	85.77	83.63	86.85

[11] Except for pretrained, NED where two of the modifications could yield a better result than all of them.

these languages, the zero-shot approach is able to outperform the prior baseline. In Basque however the results are significantly worse, with an 10 point drop in accuracy when compared to the native setting. A shallow error analysis shows that we gather few training examples for many Basque mentions. The reason is twofold. On the one hand, the number of Basque EusE dataset mentions shared with English is very low. On the other hand, Basque EusE often refers to English Wikipedia pages on local entities linked to the Basque Country, which typically are seldom mentioned in Wikipedia, and hence the training examples for these entities are scarce. In Section 6.3 we report a partial solution to these issues.

6.2. *Combining native and zero-shot*

We now analyze whether combining both native and zero-shot (from English) approaches improves the results. We train both the word expert model itself and the pretrained NED encoder for context representation[12] using both native and English data together, so we gather more training instances for training the models. Table 4 shows that the combination improves the results only marginally compared to the native approach, which indicates that incorporating English examples does not provide complementary information to the native system.

6.3. *Using non-English pivot languages*

Given a target KB in English, the most natural choice to train a zero-shot XNED system is to use English examples, but, as mentioned in Section 3.3 a pivot language other than English can be used. In fact, factors such as the typological and grammatical similarities between languages, as well as the cultural and political ties between countries, may influence the choice of the pivot language. In this section we conduct zero-shot XNED experiments for Basque, Spanish and Chinese, with English as target, when using a third language as pivot for training.

To obtain the training data in the pivot language, we map entities between the native and pivot languages by crossing the alias tables from both languages, which are both linked to English entities, via inter language Wikipedia links. Once the new alias table is built, examples in the pivot language are used to train the word expert.

The results are shown in Table 5, alongside the sizes of each Wikipedia.[13] The Table shows that, in general, using English training data is the best choice for zero-shot learning, which could be explained both by the fact that English is the target language, and the higher amount of training data due to the size of the English Wikipedia. The results in Chinese, for instance, correlate well with the sizes of the pivot Wikipedias, with Basque yielding the lowest results.

In the case of Basque, though, the Spanish training data yields the best results, 6 points higher than when using the larger English data. Note that Basque and Spanish are unrelated languages, which means that the good results cannot be attributed, in this case, to language typology. Rather, the fact that Spain and the Basque Country are culturally, geographically and politically close, make both Spanish and Basque Wikipedias share many entity mention strings such as city and person names. In fact, for many Basque cities and people, there are more training examples in the Spanish Wikipedia than in the English Wikipedia, which would explain the better results when training on Spanish Wikipedia instead of the larger English Wikipedia.

Table 4
Results for xWE (pretrained NED) as accuracy, when combining English and native Wikipedias (zero-shot and native rows, respectively). Results for zero-shot and native are copied from Table 3. Bold for best results per dataset and average.

	Basque	Spanish	Chinese	avg.
zero-shot	82.17	85.48	83.37	83.67
native	**92.41**	87.60	87.09	89.03
combined	**92.41**	**87.94**	**87.54**	**89.30**

Table 5
Cross-lingual zero-shot using different pivot languages for training. Size of each pivot language Wikipedia in millions of articles. Best accuracy results in each dataset in bold.

		Pivot	Size	Basque	Spanish	Chinese
zero-shot	xWE	*English*	6.1	82.17	**85.48**	**83.37**
		Basque	0.4	—	82.40	72.18
		Spanish	1.6	**88.50**	—	79.48
		Chinese	1.1	75.60	76.64	—

The results for Spanish, to some extent also support the need of explanatory factors beyond size, as the results using the smaller Basque data are 5 points above those obtained by Chinese, and only 3 points below those obtained with English. In this case, though, the cultural and political ties of Spanish-speaking countries to English and Basque speaking populations is not so unbalanced as for Basque, which, together with the much larger size of the English Wikipedia would explain the best results when training on English.

6.4. *Coverage of alias tables*

In this section we analyse the impact of cross-lingual alias tables in the performance of our system. These alias tables are built based on monolingual alias tables, extended using mention sharing and interlanguage Wikipedia links (see Section 4.1). In order to evaluate their impact, we focus on their coverage, measured as percentage of gold entities covered in each cross-lingual alias table. This coverage sets an upperbound for the accuracy of our XNED system, as xWE cannot correctly return an entity if the entity is missing from the cross-lingual alias table.

The figures with daggers report the coverage for the three native systems, which is over 90% in all cases. The rest of the results are for zero-shot versions. The best results are obtained when using English as a pivot, which obtains slighly larger coverage than the native systems, showing that interlanguage links are a robust and thorough resource. When using a third language as a pivot, the results are slightly lower, with larger drops when Chinese is involved.

When interlanguage link information is removed (last row in Table 6), the only option is to use directly the English alias table and rely on mention sharing. The drop with respect to the use of language links is significant. In the best case, the 66.37–69.52% of mention-entity pairs in the test dataset for Basque and Spanish are covered. As expected, mention sharing is lower for Chinese, and the figure drops dramatically for Chinese where only a 16.81% of the gold mention-entity pairs covered by the English Wikipedia.

Table 6
Coverage percentage of alias tables in test datasets (upperbound of xWE). All figures are for zero-shot, except those with a † for native systems. Last row for coverage when only mention sharing is used, without interlanguage links.

	Pivot	Basque	Spanish	Chinese
with lang-links	*Basque*	95.57†	88.68	84.99
	Spanish	95.32	90.80†	90.65
	Chinese	88.49	89.42	91.31†
	English	**95.58**	**91.21**	**91.52**
without lang-links	*English*	66.37	69.52	16.81

[12] Results for the other representation models (bag of words, pretrained XLM-R) are analogous.

[13] Source: http://wikistats.wmflabs.org/display.php?t=wp (accessed 20/5/2020)

7. Comparison with the state of the art

In this section we report a comparison with the state of the art in cross-lingual NED systems, comprising zero-shot, native, as well as the combined settings. As noted in the introduction, almost all so-called zero-shot methods in the literature combine native prior probabilities to obtain the final scores, and they do not report the results when priors are not used. So as to conduct a fair comparison with those systems, we also include the results of our zero-shot system when combined with prior probabilities. In our case, we used a simple back-off strategy, where the entity with highest prior probability for the target mention is returned when there are less than 10 examples to train the word expert for the mention.

The upper part of the Table 7 shows the results of the zero-shot system without using native priors. Here, we compare our system only with Upadhyay et al. (2018), which is the only paper reporting results without using priors. The table shows that the zero-shot system in Upadhyay et al. (2018) suffers a drastic drop in performance of almost 30 points when native priors are not used. Our system also suffers in the absence of priors, but the performance drop is less than 3 points. This result shows that our model very robust and performs well when native priors are not used, outperforming Upadhyay et al. (2018) in 32 and 27 points for Spanish and Chinese, respectively.

The results in Table 7 show that xWE obtains the best results overall in all the settings for both languages. Remarkably, our zero-shot system with no priors performs better than other zero-shot systems which do rely on native priors, and even better than competing native systems in Spanish. All in all, xWE sets a new state of the art result in XNED, both in the zero-shot and native settings.

Table 7
Results for TAC-KBP 2015 Spanish and Chinese datasets compared to the current state of the art in four settings. We report the most frequent prior baseline in two places, for easier comparison.Bold marks best results. † for non-typed (only entity linking) results reported on Upadhyay et al. (2018).

		Spanish	Chinese
zero-shot	Upadhyay et al. (2018)	53.50	55.90
	xWE	**85.48**	**83.31**
zero-shot with native priors	*prior baseline*	80.51	80.89
	Upadhyay et al. (2018)	80.30	83.90
	Sil et al. (2018)†	83.90	85.90
	xWE	**87.09**	**86.06**
native	*prior baseline*	80.51	80.89
	Upadhyay et al. (2018)	83.50	84.40
	Tsai and Roth (2016)	82.40	85.10
	xWE	**87.60**	**87.09**
combined (native + English)	Upadhyay et al. (2018)	83.50	84.40
	xWE	**87.94**	**87.54**

8. Conclusions and future work

In cross-lingual Information extraction entities mentioned in foreign texts need to be linked and disambiguated to the entries in the knowledge graph, which is typically in English. XNED systems are effective in doing that, but require large numbers of annotated data which is not always available for low-resource langauges. As a solution, zero-shot transfer learning approaches can train a XNED system based solely on English training data.

This work presents a high performing cross-lingual named-entity disambiguation model that surpasses the current state-of-the-art in cross-lingual NED in three possible training set-ups: native, zero-shot and joint training. Our approach builds one classifier for each mention string, which, contrary to prior work, allows the system to work well even when native priors are not available. In a realistic zero-shot learning scenario where no native annotation or priors are used for training, our system surpasses by more than 20 point the previous state-of-the-art performance in zero-shot XNED.

We experimented different multilingual transfer strategies, showing that better results are obtained with a purpose-built multilingual pre-training method compared to state-of-the-art generic multilingual models such as XLM-R. Our results also show that, surprisingly, English is not always the best language to train a XNED system into English. For instance, Spanish is more effective when training a zero-shot XNED system that disambiguates Basque mentions with respect to an English knowledge graph.

One limitation of current zero-shot XNED systems, including ours, is that they make use of an alias table for the candidate generation step, which is tipycally based on Wikipedia lang-links. This requirement sets an upper bound in the performance that can be obtained for a language, which depends on the size of the native Wikipedia. Some recent work (Wang, Lv, Lan, & Zhang, 2018; Chen, Shi, Zhou, & Roth, 2020) has shown that it is possible to induce candidates based on cross-lingual graph alignment algorithms alone, without the need of Wikipedias. We would like to explore whether such candidate generation techniques, in combination with our XNED system, allow to perform XNED languages for languages which do not have Wikipedias or have small Wikipedias.

The hand-labeled dataset for Basque-English Cross-lingual evaluation, the unsupervised cross-lingual word embeddings and all the pre-trained models in this paper, as well as the code to reproduce results is available for reproducibility.[14]

CRediT authorship contribution statement

Ander Barrena: Conceptualization, Methodology, Software, Investigation, Resources, Writing - original draft, Funding acquisition. **Aitor Soroa:** Validation, Writing - review & editing, Supervision. **Eneko Agirre:** Validation, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially funded by the Basque Government (IXA excellence research group (IT1343-19) and DeepText project), Project BigKnowledge (Ayudas Fundación BBVA a equipos de investigación científica 2018) and via the IARPA BETTER Program contract 2019–19051600006 (ODNI, IARPA activity). Ander Barrena enjoys a post-doctoral grant ESPDOC18/101 from the UPV/EHU and also acknowledges the support of the NVIDIA Corporation with the donation of a Titan V GPU used for this research. The author thankfully acknowledges the computer resources at CTE-Power9 + V100 and technical support provided by Barcelona Supercomputing Center (RES-IM-2020–1-0020).

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Agirre, E., & Edmonds, P. (2007). *Word Sense Disambiguation: Algorithms and Applications* (1st ed.). Incorporated: Springer Publishing Company.

[14] https://github.com/anderbarrena/xNED

Artetxe, M., Labaka, G., & Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)* (pp. 5012–5019).

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 789–798).

Barrena, A., Soroa, A., & Agirre, E. (2018). Learning text representations for 500k classification tasks on named entity disambiguation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 171–180). Association for Computational Linguistics.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Journal of web semantics, 7*, 154–165.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Chang, P.-C., Galley, M., & Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation* (pp. 224–232).

Chen, M., Shi, W., Zhou, B., & Roth, D. (2020). Cross-lingual entity alignment for knowledge graphs with incidental supervision from free text. arXiv preprint arXiv: 2005.00171.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Conneau, A., & Lample, G. (2019). *Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems* (pp. 7057–7067). Curran Associates Inc..

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management, 51*, 32–49. https://doi.org/10.1016/j.ipm.2014.10.006. URL: http://www.sciencedirect.com/science/article/pii/S0306457314001034.

Fernandez, I. (2012). *Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa. Ph.D. thesis Lengoaia eta Sistema Informatikoak Saila*. UPV/EHU University of the Basque Country.

Ji, H., Grishman, R., & Dang, H. (2011). Overview of the text analysis conference TAC2011 Knowledge Base Population Track. In *TAC 2011 Proceedings Papers. National Institute of Standards and Technology (NIST)*.

Ji, H., Nothman, J., Hachey, B., & Florian, R. (2015). Overview of text analysis conference TAC2015 Knowledge Base Population Track, tri-lingual entity discovery and linking. In *TAC 2015 Proceeding Papers. National Institute of Standards and Technology (NIST)*.

Kasai, J., Qian, K., Gurajada, S., Li, Y., & Popa, L. (2019). Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Logeswaran, L., Chang, M.-W., Lee, K., Toutanova, K., Devlin, J., & Lee, H. (2019). Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. Academic Press.

McNamee, P., Mayfield, J., Lawrie, D., Oard, D., & Doermann, D. (2011). Cross-language entity linking. In Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 255–263). Asian Federation of Natural Language Processing.

Mikolov, T., Le, Q.V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250.

Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016). From freebase to wikidata: The great migration. In Proceedings of the 25th International Conference on World Wide Web WWW '16 (p. 1419–1428). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. URL: https://doi.org/10.1145/2872427.2874809. doi: 10.1145/2872427.2874809.

Rijhwani, S., Xie, J., Neubig, G., & Carbonell, J. (2019). Zero-shot neural transfer for cross-lingual entity linking. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 6924–6931). volume 33.

Ruder, S. (2019). Neural Transfer Learning for Natural Language Processing. Ph.D. thesis National University of Ireland, Galway.

Sil, A., Kundu, G., Florian, R., & Hamza, W. (2018). Neural cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Tsai, C.-T., & Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 589–598). Association for Computational Linguistics.

Upadhyay, S., Gupta, N., & Roth, D. (2018). Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2486–2495). Association for Computational Linguistics.

Veloso, A., Ferreira, A. A., Gonçalves, M. A., Laender, A. H., & Meira, W. (2012). Cost-effective on-demand associative author name disambiguation. *Information Processing & Management, 48*, 680–697. https://doi.org/10.1016/j.ipm.2011.08.005. URL: http://www.sciencedirect.com/science/article/pii/S0306457311000847.

Wang, Z., Lv, Q., Lan, X., & Zhang, Y. (2018). Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 349–357).

Zhao, G., Wu, J., Wang, D., & Li, T. (2016). Entity disambiguation to wikipedia using collective ranking. *Information Processing & Management, 52*, 1247–1257. https://doi.org/10.1016/j.ipm.2016.06.002. URL: http://www.sciencedirect.com/science/article/pii/S0306457316301893.

Zhou, S., Rijhwani, S., & Neubig, G. (2019). Towards zero-resource cross-lingual entity linking. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) (pp. 243–252). Association for Computational Linguistics.

Zhou, S., Rijhwani, S., Wieting, J., Carbonell, J., & Neubig, G. (2020). Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics, 8*, 109–124.