Crosslingual Section Title Alignment in Wikipedia

Djellel Difallah
NYU Abu Dhabi
Abu Dhabi, UAE
djellel@nyu.edu

Diego Saez-Trumper
Wikimedia Foundation
Barcelona, Spain
diego@wikimedia.org

Eriq Augustine
UC. Santa Cruz
California, USA
eaugusti@ucsc.edu

Robert West
EPFL
Lausanne, Switzerland
robert.west@epfl.ch

Leila Zia
Wikimedia Foundation
California, USA
lzia@wikimedia.org

***Abstract*—Sections are the building blocks of Wikipedia articles. They are used by editors to create a structure for the content of articles, which in turn improves reading and editing workflows. Today, millions of carefully curated *section titles* exist in more than 160 actively edited Wikipedia languages as standalone components of a larger system. Understanding the connection and correspondence of section titles across languages presents various application opportunities such as article template recommendation, i.e., given a source language article, we can generate a skeleton of section titles for a target language. Inspired by this real-world data mining problem, the present paper introduces the problem of aligning section titles across Wikipedia languages and proposes a probabilistic method for identifying such correspondences. Instead of applying translation tools to section titles (which may generate out-of lexicon titles), we develop a supervised model that identifies cross-language mappings based on section content features. We collected a ground-truth dataset created for this purpose with the help of volunteers. In addition, we use Probabilistic Soft Logic to model the dependencies between multilingual section pairings. We show that our approach performs better than machine translation solutions in about 80% of the language pairs, including distant language mappings such as Arabic to Russian or French to Japanese and in many of the more closely related languages such as French to Spanish.**

***Index Terms*—Instance Matching; Cross-lingual Alignment; Probabilistic Soft Logic; Wikipedia; Crowdsourcing**

I. Introduction

With more than 45 million articles across roughly 160 actively edited languages, Wikipedia is the largest encyclopedia ever built. It is edited more than 15 million times a month by hundreds of thousands of volunteers [1] across the globe who contribute content in their local languages about the topics they choose to write about and following the local norms and guidelines of their Wikipedia language edition. While the sum of knowledge in Wikipedia is vast, the diversity of forms and languages in which Wikipedia is written in often creates barriers for the accessibility of this vast body of knowledge [2] and Wikipedia volunteer editors and readers are those who are directly affected by it.

Today, the knowledge in Wikipedia is unevenly distributed over the languages the content is written in. More than 25% of Wikipedia articles appear in only one language and only 10% of Wikipedia languages have more than 1 million articles.[1] These statistics only surface the tip of the iceberg: the content within the existing Wikipedia articles can be drastically different when considering the breadth and depth of the content represented. In English Wikipedia alone, more than 37% of the articles are flagged as stubs, i.e., short articles. The isolation of content in Wikipedia within the different editions creates numerous challenges for Wikipedia as a platform. One such challenge is that for languages with very few editors but a large pool of (monolingual) readers, the large-scale creation and maintenance of content may not be possible due to resource constraints. As a result, the millions of people who visit such Wikipedia languages cannot have access to knowledge that may be already available in some of the other Wikipedia languages. For example, at the time of writing UNIVERSE does not have an article in Malagasy, a language spoken by 25 million people. On top of this, Wikipedia editors contributing to a Wikipedia language may not be aware that similar content exists in one or more of the other languages. This can result in duplication of effort and inefficiencies that could be addressed if the platform is aware of the availability of content in other languages.

Given the importance of the connectivity and flow of knowledge across Wikipedia languages, there have been multiple attempts to break the language barriers in Wikipedia. These attempts can be divided into two categories. The first set of approaches focus on entity-level alignment of Wikipedia articles. The most prominent example is the inter-language links panel that connects an article in a given language to its equivalent articles in other languages. Another effort is Wikidata [3], which provides a unique and multilingual resource for structured and machine readable, entity representation. The entity-based approaches have two main challenges: the bridges they create are at the article level which does not provide enough information about the body of the article, and the connections created can be imprecise [2], [4]. The second set of approaches aim at finding word, or sentence, level alignment [5], [6]. The main downside of these approaches is that the methods used often do not scale across languages or the results become uninterpretable due to the level of granularity in the outputs.

In this paper, we propose Wikipedia sections as the intermediate level of granularity for aligning knowledge across Wikipedia languages. We introduce the problem of section title alignment across Wikipedia languages. Specifically, we aim at finding the equivalent section titles for a given pair

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

of languages. While this can be categorized as a translation problem, in practice (1) we need to adhere to the existing lexicon of titles adopted in a given Wikipedia language version, and (2) direct title translation is context-free and may yield literal or irrelevant alignments. Hence, we focus our efforts on building a crosslingual translation detection classifier system. First, we use section content features (cross-lingual word embeddings) to represent a given section-title. Next, we train a translation classifier to detect if a pair of titles from two different languages are equivalent, and a synonym-classifier to detect if a pair of titles from the same language are synonymous. Finally, we use Probabilistic Soft Logic [7], a machine learning framework for performing structured prediction over probabilistic graphical models, to model the interdependence of sections in the translation and synonym classifiers. We show that using PSL and including both translation and synonym classifiers, our section title alignment model performs better in roughly 80% of the language pairs compared to the baseline, which is the output of a prominent proprietary translation service.

Use-cases and Relevance. The alignment model developed in this research helps us understand the large multilingual corpus of the section-title lexicon used across hundreds of languages. It also opens opportunities for a variety of applications in Wikipedia. We name a few here. (i) Article Editing: The model will empower a recommender systems for article expansion; it uses a reference article from a source language to generate relevant sections to be added in a target language. (ii) Translation: The same model can also be used to recommend the relevant sections to translate for Wikipedia contributors who are engaged in this purposeful activity. (iii) Crosslingual representation: More generally, the section title alignment model will benefit any system that attempts to represent Wikipedia articles in a language-independent manner, a key step for scaling many of the language-specific machine learning solutions developed today to all Wikipedia languages.

Contributions. Our main contributions are as follows:

- We introduce the problem of section title alignment across Wikipedia languages. (Sec. III).
- We create a section title alignment model using local and global features. We build a crosslingual translation detection classifier and an intralingual synonym detection classifier. We use Probabilistic Soft Logic (PSL) to model the dependencies of Wikipedia sections. (Sec. IV)
- We evaluate our models based on a large multilingual labeled dataset and compare the models with proprietary machine translation services, showing that our model performs better in roughly 80% of the language pairs, even in rare language pairs. (Sec. V–VI)

Project repository. All code, data, and results from this research are shared publicly. This includes the release of a dataset containing multilingual dictionaries generated by experienced Wikipedia editors as well as crosslingual vector representations of sections.[2] The section title alignments are made available through a public API. [3]

II. Related Work

To develop a model for aligning Wikipedia sections across languages, our study draws on three main different lines of research, described as follows.

Automatic translation and crosslingual word embeddings. In recent years, the research and development on automatic machine translation has transitioned from statistical machine translation models [8] to deep learning based models [9]. However, both approaches often require identical parallel corpora to learn from, a limiting factor for including many language pairs in the model development and application stage. To address the parallel corpora challenge, a different approach was developed where first the embeddings were separately trained for each language and then a linear transformation aligned them in the same vector space [10]. This approach inspired several studies and implementations [11]–[14]. In this paper, we build on the earlier research of FastText alignments where the authors use a proprietary multilingual dictionary to create word alignments in Wikipedia and without the need for parallel corpora [15]. We address the need for proprietary multilingual dictionaries that are usually difficult to have access to and are often small by proposing a Wikidata-based solution (Sec. IV).

Probabilistic Soft Logic. Probabilistic Soft Logic (PSL) is a machine learning framework for creating and inferring over hinge-loss Markov random fields (HL-MRFs), providing an easy syntax based on first-order logic [16]. Among other applications, PSL has been used for entity resolution [17], knowledge graph construction [18], online inference [19], learning latent variables [7], and topic modeling [20]. In the present work, we use PSL to model the interdependence across language alignments, allowing a multilingual approach to address the section title alignment problem instead of using isolated languages pairs.

Wikipedia knowledge alignment. The research on the alignment of knowledge in Wikipedia is divided into two parts: studies that aim to align Wikipedia articles as concepts [3], [21] and those that focus on the word, sentence, or paragraph alignment across languages and projects [5], [6], [22]. Both these approaches provide alignments that are either too granular and not easily interpretable, or too high level. In the latter case, they can also suffer from imprecisions introduced by the article-as-concept assumption [2], [4]. We propose to create alignments using sections as the intermediate level of granularity.

Instance Matching. Our task is related to the problem of instance matching which is defined as the identification of a similar real-world object present in independent datasets. In particular, the alignment of instances in Linked Open

[2] https://github.com/digitalTranshumant/wmf-interlanguage

[3] The following call generates the correspondence of the English section title CRITICISM to Spanish: https://secrec.wmflabs.org/API/alignment/en/es/Criticism

Data (LOD) datasets aims to link similar instances with an `owl:sameAs` link. With only two datasets to align, the number of candidate matching pairs grows quadratically with the size of the data, making the matching task intractable for large datasets. Several methods based on blocking [23], large-scale crowdsourcing with probabilistic reasoning [24], [25], and graph embeddings [26], [27] have been proposed. Although our problem is formulated differently, we leverage similar techniques to tackle the computational complexity that arises from aligning section titles from hundreds of languages using machine learning, PSL, and crowdsourcing.

III. THE SECTION TITLE ALIGNMENT PROBLEM

In this section we define some key terminology and notation and present the section-title alignment problem. We support our description with the illustration in Fig. 1.

We denote Wikipedia language editions with upper case letters as L, M, N, and section titles with lower case letters as s, t, u. When we intend to emphasize that a section title s is from a language L, we add L as a superscript, writing s^L. We distinguish between individual *section instances* that refer to the actual textual content of a given section, and *section titles* such as AWARDS, BIOGRAPHY, GEOGRAPHY. As such, in our data model, a section title may have multiple section instances, while a section instance has a unique section title.

Abstractly, the section title alignment problem may be modeled as a link prediction problem in an n-partite graph, where n is the number of languages, the nodes of the graph represent section titles, and the links represent the crosslingual alignment. In such a graph, corresponding section titles across different languages will create dense neighborhoods in the graph.

In practice, we treat the section title alignment problem as a retrieval problem. Given two Wikipedia languages L, M and a section title s^L, we must rank all section titles of language M such that the section titles corresponding to s^L appear at the top of the ranking, *e.g.*, if L is German and M is English, and $s^L =$ AUSZEICHNUNGEN, then AWARDS should be ranked at the top of the ranked list of section titles in M.

Note that in the above we used the plural "the section titles corresponding to s^L" as within the same language, section titles may be fully or partially synonymous (*e.g.*, HONORS and HONORS AND AWARDS; LIFE and EARLY LIFE). As a result, the same section title in one language may correspond to several section titles in another language. Also, given the knowledge gaps in Wikipedia's content [28], not all section titles have corresponding sections in all other languages. Finally, it is important to note that the notion of "correspondence" is not as strict as direct translation. Instead it captures encyclopedic equivalence, considering the different conventions and usages in each Wikipedia language community. For instance, the section CURIOSIDADES in Spanish corresponds to TRIVIA in English while one is not the literal translation of the other.[4]

The above observations make it difficult to formally define the notion of section title correspondence; instead, we take a data-driven, human-centric approach, whereby we present Wikipedia contributors with section titles in language L and ask them to provide the corresponding section title (or titles) from language M. The goal of the section title alignment model, therefore, is to generalize these labels to all unlabeled pairs of sections in the two languages.

IV. PROPOSED SOLUTION

In this section, we propose a solution to address the problem of section title alignment across Wikipedia languages. Our solution can be broadly organized into two levels, local and global, depending on the type of features and algorithms utilized.

Local Translation Classifier (LTC). We design a crosslingual translation detection classifier, using section content features. This classifier receives information about a pair (s, t) of section titles as input and outputs an estimate of the probability that s translates to t. The translation classifier works with features of s and t, including the section title strings and the database of all section instances that appeared under each.

The local level, however, misses out on two important information about the section titles: (1) section titles within the same language can be synonymous, and (2) section titles are not isolated and can have dependencies that need to be captured as part of any model that aims to align them. Fig. 1 shows some of these section title dependencies. For instance, if EN:AWARDS AND RECOGNITION is a translation of ru:Награды и премии, and ru:Награды и премии is a translation of FR:RÉCOMPENSES ET DISTINCTIONS, by transitivity, this increases the probability that EN:AWARDS AND RECOGNITION is also a translation of FR:RÉCOMPENSES ET DISTINCTIONS. Analogously, the transitive property can be applied to synonyms. Finally, if FR:RÉCOMPENSES ET DISTINCTIONS is a translation of EN:AWARDS AND RECOGNITION, and EN:AWARDS AND RECOGNITION is synonymous with EN:AWARDS AND HONORS, then this increases the probability that FR:RÉCOMPENSES ET DISTINCTIONS is also a translation of EN:AWARDS AND HONORS.

Global Translation Classifier (GTC). For modeling such dependencies at the global level, an intralingual synonym detection model and joint inference methods are required. We build the synonym detection classifier using gradient-boosted trees. We then use use *Probabilistic Soft Logic* (PSL) [16], which offers an expressive language for specifying section dependencies, as well as efficient learning and inference algorithms. PSL works on a graph where section titles are nodes, and edges between section titles are initially weighted with the translation probabilities which are the outputs of the local classifier. As the result of the global, joint inference, the edge weights are updated such that they more closely follow the aforementioned rules (such as translation transitivity).

In the remainder of this section, we first describe the local translation classifier features in details (Sec. IV-A). We then

[4] Our exposition will at times be more intuitive when phrased in terms of "translation" rather than "alignment", but the reader should keep the above caveat in mind.

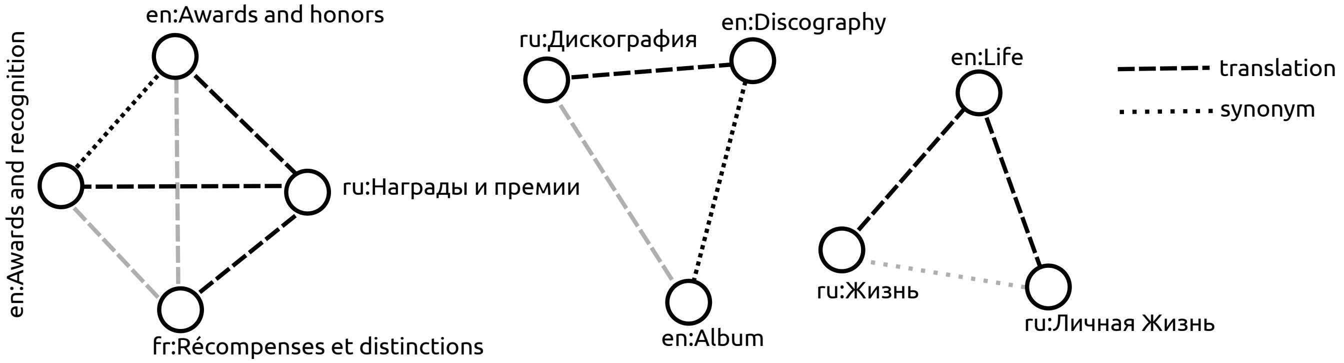


Figure 1. Ideal crosslingual section title alignment for three languages. Weights for all alignment and synonymy edges are individually predicted by the respective local alignment (Sec. IV-A) and synonymy (Sec. IV-B) classifiers. Grey edges indicate a low prediction score from the local classifier. To improve on the local predictions, we feed them into a global alignment model based on Probabilistic Soft Logic (Sec. IV-B), which encourages edges to observe intuitive rules such as transitivity as seen in the grey edges in the leftmost clique. These edges get low values from local alignment classifier, but can be inferred via transitivity of alignment relation. The grey edges in the center and rightmost cliques can similarly be inferred in the global alignment model.

present the global translation classifier based on our PSL model with and without synonym detection (Sec. IV-B).

A. Local translation classifier

The input of the local translation classifier is a pair (s^L, t^M) of section titles, from languages L and M, respectively. The output of the classifier is the estimate of the probability that s^L translates to t^M. The pair of section titles is represented as a vector of features extracted from (1) the section titles themselves, (2) the content of the sections with the respective titles, and (3) the co-occurrence of the two section titles in articles that, although written in different languages, are about the same concept. The classifier is based on gradient-boosted trees [29], trained on a hand-labeled ground truth collected for the purpose of this study (Sec. VI).

To evaluate the classifier in the rankedretrieval framework delineated in Sec. III, we use output probabilities as scores for ranking. The performance of the classifier heavily depends on the features used, and identifying powerful features is indeed one of our contributions. Hence, the remainder of this subsection is dedicated to describing the abovelisted features in more detail.

Section-title features. The first set of features is extracted from the section titles themselves. In closely related languages, corresponding section titles are frequently similar or even identical on the surface (*e.g.*, Spanish VIDA and Italian VITA); additionally, certain section titles (*e.g.*, decades or proper nouns) are identical even for many of the unrelated languages. To capture this effect, we include the Levenshtein edit distance between s^L and t^M as a feature.

To capture semantic similarities between section titles that are not immediately similar or identical we use crosslingual vector representations of the two section titles using FastText [15].[5] More specifically, we represent words as numerical vectors in a 300-dimensional space such that semantically similar words have similar vectors across languages. As section titles may consist of multiple words, the vector for a section title is computed as the mean of the crosslingual word vectors of the constituent words. Two section titles are then compared via their cosine similarity. Note that for aligning word vectors across languages, Smith *et al.* [15] use words that might appear identically across multiple languages (*e.g.*, proper nouns) as anchors. Unfortunately, this approach fails to align languages with different scripts. Hence, we modify their approach as follows. We first map words to Wikipedia article titles via exact string matching; since every Wikipedia article is linked to a language-independent concept in the Wikidata knowledge base, this establishes a (partial) mapping from words to language-independent concepts, and we may use words from different languages that map to the same concept as alignment anchors.

Section-content features. Two section titles s^L and t^M from different languages are more likely to be translations of one another if sections from language L with title s^L cover similar items as sections from language M with title t^M. To capture this intuition, we design features capturing the similarity between the content of sections. We represent each individual section instance as a crosslingual embedding vector via the IDF-weighted average of the embeddings of the words appearing in the section [14]. For each concept c such that s^L appears in L's version of c, and t^M appears in M's version of c, we compute the cosine similarity of the embedding vectors corresponding to the instances of s^L and t^M, respectively, and then aggregate (via mean, median, and sum) the cosine similarities across all concepts c.

We also compute the following properties for each section instance, aggregated (via mean and median) over all section instances with the same title, and use the difference between the aggregate values for s^L and t^M as features: length (in terms of the absolute number of characters, as well as a ratio with respect to the length of the entire article); number of links; links density (the number of links divided by section length); and position in the article (*e.g.*, INTRODUCTION

[5] https://github.com/Babylonpartners/fastText_multilingual

usually appears at the beginning, and SEE ALSO, at the end).

Finally, we compute the ranks of s^L and t^M in terms of the frequency of occurrence across Wikipedia in languages L and M, respectively, and add the difference as a feature.

Co-occurrence features. The final set of features is built based on the intuition that section titles from different languages that tend to co-occur in articles about the same concepts are more likely to be translations of one another. For example, if a section titled DISCOGRAPHY occurs in an English article, then its translation DISCOGRAFÍA tends to occur in the Spanish version of the same article. On the other hand, if two section titles never co-occur in articles about the same concept, they are unlikely to be translations. Following this intuition, given a pair (s^L, t^M) of section titles, we define the concept co-occurrence feature as the number of Wikidata concepts for which s^L occurs in L's version and t^M in M's version.

Since some section titles (*e.g.*, REFERENCES, SEE ALSO) occur widely across all articles, taking a role similar to that of stop words in language modeling, we also include an IDF-weighted version of the concept co-occurrence feature (weighting a section title that appears in k out of all N arts by a factor of $\log(N/k)$). Note that this approach helps with filtering out all the sections that never co-occur across languages, significantly reducing the number of candidates for alignments.

B. Global translation model

In this section, we build a global alignment model by addressing two challenges faced by the local translation classifier introduced above. First, we build a synonym detection classifier to allow for the identification of synonym sections in a given language. Second, we model the different types of dependencies that exist between sections.

Synonymy classifier. To predict if two section titles s and t from the same language L are synonymous (*e.g.*, EXTENDED PLAYS and EPS), we adopt a similar approach as we use for predicting whether two section titles from different languages are translations of each other. Note that the synonymy classifier is built only to feed into the PSL model and is not intended as a standalone model.

We use section-title features (Sec. IV-A) with monolingual (in place of crosslingual) embeddings. We also add a binary feature indicating whether s is a substring of t or vice versa. The latter feature aims to capture overlaps such as AWARDS and HONORS AND AWARDS. Additionally, we add co-occurrence features designed specifically to detect synonymy. The premise here is that two synonyms tend to appear in the same context, but should (almost) never appear together in the same article (*e.g.*, AWARDS and HONORS AND AWARDS both tend to co-occur with CAREER, but AWARDS nearly never co-occurs with HONORS AND AWARDS). We hence add two features: (1) To capture co-occurrence frequency, we simply count the number of articles in language L that contain both section titles. (2) To capture contextual similarity, we represent each section title's context as an IDF-weighted vector of the sections it co-occurs with and compute the cosine similarity of the two contexts.

Probabilistic Soft Logic. Supervised learning algorithms using local features, such as those introduced above for detecting translation, treat all section pairs as isolated items; but they are not. On the contrary, there are clear dependencies, as demonstrated in the beginning of Sec. IV. To capture those dependencies, we must jointly perform inference over all pairs, and we turn to Probabilistic Soft Logic (PSL) [16] for doing that.

PSL allows for the specification of weighted logical rules that can encode domain knowledge about the model. The variables in these rules take continuous values in $[0, 1]$, and the rules themselves have a continuous penalty in $[0, 1]$ with a fully satisfied rule taking a penalty of 0.0 and a fully unsatisfied rule taking a penalty of 1.0. The higher the weight of a rule, the higher the penalty for violating it. Additionally, the weights need not be specified by hand, but can be learned from (partially) labeled data in a supervised fashion.

For every section pair (s^L, t^M) where $L \neq M$, we want to infer whether s^L is a translation of t^M. For this, we introduce a variable $\mathrm{Tr}(s^L, t^M)$ to the model. Similarly, for every (s^L, t^L), we want to infer if s^L is synonym of t^L and we introduce a variable $\mathrm{Syn}(s^L, t^L)$ to the model. Using these variables, we can apply the following rules:

Transitivity. If s^L is a translation of t^M and t^M is the translation of u^N, then s^L should be the translation of u^N. Formally,

$$\mathrm{Tr}(s^L, t^M) \ \& \ \mathrm{Tr}(t^M, u^N) \ \rightarrow \ \mathrm{Tr}(s^L, u^N).$$

Similarly, transitivity should hold for the synonym relation:

$$\mathrm{Syn}(s^L, t^L) \ \& \ \mathrm{Syn}(t^L, u^L) \ \rightarrow \ \mathrm{Syn}(s^L, u^L).$$

Symmetry. Both translation and synonymy are symmetric relations:

$$\mathrm{Tr}(s^L, t^M) \ \rightarrow \ \mathrm{Tr}(t^M, s^L)$$

$$\mathrm{Syn}(s^L, t^L) \ \rightarrow \ \mathrm{Syn}(t^L, s^L)$$

Consistency. If s^L and t^L are synonymous in language L, and s^L is a translation of u^M , then t^L is also a translation of u^M:

$$\mathrm{Syn}(s^L, t^L) \ \& \ \mathrm{Tr}(s^L, u^M) \ \rightarrow \ \mathrm{Tr}(t^L, u^M)$$

The above rules implement soft constraints between the inferred values only. We have yet to specify that the inferred values should be close to the values predicted by the local translation and synonymy classifiers. We achieve this via the following rules (Tr_0 and Syn_0 specify a constant for every pair of section titles output from the local classifiers):

$$\mathrm{Tr}_0(s^L, t^M) \ \rightarrow \ \mathrm{Tr}(s^L, t^M)$$

$$\mathrm{Syn}_0(s^L, t^L) \ \rightarrow \ \mathrm{Syn}(s^L, t^L)$$

Table I
FOR EACH LANGUAGE WE SELECT THE TOP RANKED SECTIONS THAT CUMULATIVELY ARE USED IN AT LEAST 75% OF THE TOTAL NUMBER OF ARTICLES IN THE LANGUAGE. THESE ARE THE SECTIONS TO BE TRANSLATED ("SOURCE" SECTIONS). AS CANDIDATES FOR TRANSLATION, WE CONSIDER ALL THE SECTIONS TITLES APPEARING IN THE TARGET LANGUAGE ("TARGET" SECTIONS).

Language	Source Sections	Target Sections
ar	981	203,826
en	999	1,697,060
es	527	436,460
fr	529	565,292
ja	759	448,917
ru	485	413,502

Table II
THE NUMBER OF LABELS PER (DIRECTED) LANGUAGE PAIR. LANGUAGE PAIRS WITH LESS THAN TEN LABELS WERE DISCARDED.

To From	ar	en	es	fr	ja	ru
ar	-	382	59	136	-	23
en	213	-	568	668	380	643
es	-	359	-	335	-	16
fr	30	342	100	-	60	43
ja	-	68	-	47	-	-
ru	-	295	15	207	33	-

Computational considerations. The large number of variables imposes a heavy computational burden. We therefore make the graph more sparse by only including a translation edge $\mathrm{Tr}(s^L, t^M)$ if there exists at least one Wikidata concept c such that s^L appears in language L's version of c, and t^M appears in language M's version of c. Limitations and possible improvements for this decision are discussed in Sec. VII.

V. DATASETS

Six languages were selected for this study considering the diversity in terms of scripts (Cyrillic, Latin, etc.) and the family (Indo-European, Afro-Semitic, etc.). These languages are Arabic, English, French, Japanese, Russian, and Spanish. We describe the utilized datasets, baseline and preprocessing steps below.

Wikipedia and Wikidata extractions. In order to extract Wikipedia article sections and compute some of the features described in Sec. IV, we utilized the public Wikipedia dump[6] of Arabic, English, French, Japanese, Russian and Spanish Wikipedia released on 2018-08-08. We extracted all section titles, and ranked them by frequency of occurrence in every language. We then defined a threshold for the ranked list and only kept the top ranked section titles that cumulatively represented 75% of the articles in that language (see Table I). Note that this step is required to limit the number of sections that need translations due to the fact that collecting ground truth labels from experienced Wikipedia editors is costly with respect to their volunteer time. We also know that section frequencies follow a power law distribution. Also note that while only a subset of sections in every (source) language were considered, all sections of the target languages were considered as potential section translation candidates (see the right-hand-side column in Table I).

We also utilized Wikimedia's SQL replicas[7] to obtain the Wikidata item corresponding to each of the articles extracted in the previous step. Wikidata items are crosslingual concept level IDs, which allow — among other things — identification of two articles about the same concept in different languages. For example, the English Wikipedia page about *Artificial Intelligence* corresponds to the Wikidata item Q11660, which is the same for *Inteligencia Artificial* in Spanish Wikipedia and 人工知能 in Japanese Wikipedia.

A. Crowdsourced Mappings

For this study, we needed to build two labeled datasets for (i) the translation task, and (ii) the synonymy task. The ground truth of whether a section title is a translation or synonym for another section title needed to be collected from experienced Wikipedia editors since as described in Sec. III the synonymy and translation relations are of the correspondence nature and encyclopedic expertise is required to be able to make the judgment on such relations. To locate such expertise among the volunteers, we used the MediaWiki Babel extension[8] that lists the language proficiency of the editors. Out of all users who had Babel information available, we requested a contribution from those who had an advanced or higher level competence in a given language.

Translation Task. The labels for the translation task were obtained via a web application depicted in Fig. 2. The application allows users to select the source and target language given their expertise. The users were also given suggestions for the section title in the target language. These suggestions are pulled from a ranked list of section titles in the target language. By default, users are required to add at least one mapping per section title, but there was no upper bound on the number of mapping they can add. Once a mapping for a section title is obtained, it is marked as done, and not shown to other users. While for some pairs of languages, such as Spanish to English we have successfully translated all the section titles, for more unusual pairs such as Japanese to French we have obtained a subset of translations. All the language pairs for which we have less than ten translations have been discarded (see Table II).

Synonym Task. For gathering synonyms, we created a spreadsheet for each language and asked users to mark whether two section titles are identical, overlapping, or different. We showed one thousand pairs of sections to each user, from a stratified sample (based on the word embedding distance). [9]

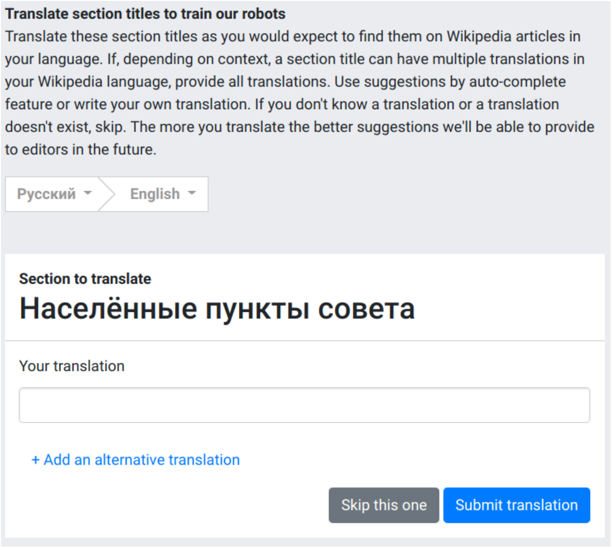
Baseline datasets. Since our main focus is on section title alignment, we use as a baseline a commercial translation

[6] https://dumps.wikimedia.org

[7] https://quarry.wmflabs.org

[8] https://www.mediawiki.org/wiki/Extension:Babel

[9] The full log of the data collection from the editors is publicly available: https://phabricator.wikimedia.org/T195001



Translate section titles to train our robots

Translate these section titles as you would expect to find them on Wikipedia articles in your language. If, depending on context, a section title can have multiple translations in your Wikipedia language, provide all translations. Use suggestions by auto-complete feature or write your own translation. If you don't know a translation or a translation doesn't exist, skip. The more you translate the better suggestions we'll be able to provide to editors in the future.

Section to translate

Населённые пункты совета

Your translation

Figure 2. Screenshot of the web application interface for labeling section title pairs. The application allows users to select the source and target language and shows a section title to translate. The editors can enter one or multiple relevant section titles in the target language.

service (Google Translate) that exposes a translation API for hundreds of languages. Each section was translated to the other five languages and this translation was compared with the ground-truth labels.

VI. Evaluation

Here we describe our results considering the section title alignment problem defined in Sec. IV. For the translation tasks, we first compare our results against a Multi-lingual Translation service – We use Google Translate. In this case only precision @1 is compared since the API used for obtaining translation results provides the single best translation option. We then provide results for the Global Classifier without and with synonym detection classifier used. In these cases, we provide precision @1, @3, and @5.

Given that our system aims to scale to any language pairs, we propose a methodology that can be trained without the availability of labels in a given pair. To this end, we split the data by language, train our models considering $k-1$ languages, and test on the k-th language.

Local translation classifier. Table III shows that the local translation classifier (LTC) beats out the Multi-lingual Translation service (MLT) in 18 of the 23 languages pairs. Major improvements ($\geq$ 0.50) were seen in four (17%) languages pairs, and significant improvements ($\geq$ 0.10) in fifteen language pairs (65%). There is a single case with the same performance and just four language pairs where the MLT performs better. However, the difference in performance for these language pairs is always less than 0.07.

Table III

Performance on the translation task. The numbers in the table indicate the precision@1 for the Multi-lingual Translation (MLT), the Local Translation Classifier (LTC), the Global Translation Classifier without synonyms (GTC w/o Syn) and the Global Translation Classifier with synonyms (GTC). The highest scoring method for each language pair is in bold.

Language Pair	MLT	LTC	GTC w/o syn	GTC
ar-ru	0.09	0.78	0.57	**0.83**
ar-en	0.35	0.73	0.72	**0.74**
ar-es	0.58	**0.63**	0.42	**0.63**
ar-fr	0.45	**0.57**	0.50	**0.57**
en-fr	0.60	0.76	0.66	**0.79**
en-ja	0.55	0.52	0.51	**0.56**
en-ru	0.38	0.70	0.64	**0.71**
en-ar	0.38	0.53	0.46	**0.55**
en-es	**0.83**	0.80	0.73	**0.83**
es-ru	0.19	0.69	0.25	**0.69**
es-en	0.71	0.77	0.63	**0.77**
es-fr	0.59	0.80	0.62	**0.83**
fr-ja	0.32	0.38	0.35	**0.40**
fr-ar	0.17	0.37	**0.50**	0.40
fr-ru	0.12	**0.63**	0.43	**0.63**
fr-en	0.58	0.77	0.59	**0.78**
fr-es	0.18	0.69	0.45	**0.79**
ja-en	0.21	0.47	0.32	**0.49**
ja-fr	**0.34**	0.28	0.23	0.30
ru-en	0.58	0.68	0.57	**0.71**
ru-es	**0.67**	**0.67**	0.33	**0.67**
ru-fr	0.42	**0.53**	0.42	0.52
ru-ja	**0.30**	0.24	0.27	0.24

Table IV

The most important features (Gain score) for the local translation classifier.

Feature	Importance
Distance Embedding WikiData	.120
Links Similarity (mean)	.072
Rank Frequency Distance	.045
Co-occurence LangTo-TfIdf (mean)	.057
Distance Embedding pre-trained	.044
Co-occurence LangTo-TfIdf * Freq	.046
Edit Distance	.044
Links Similarty (Sum)	.046
Position Distance (mean)	.031

Looking at Table IV, we see that the most relevant feature for our classifier is the Wikidata-based aligned embedding, followed by the link similarity, and the co-occurrence count. It is interesting to note that although the Wikidata-based aligned embeddings are clearly the most important feature, there is no feature contributing more than 12%, suggesting that there are many features with high predictive power.

Global translation classifier without synonymy detection (GTC w/o syn). Next, we describe our results considering the global section title alignment problem defined in Sec. IV-B. We first share the result of the global translation classifier without synonymy detection. Table III shows the precision @1 for the global translation classifier (GTC) along with the local translation classifier (LTC) against a baseline, Multi-lingual Translation (MLT), the most immediately available solution in the absence of a dedicated model for section translation. We observe that in all but two language pairs, the global translation classifier performed equally well or better than an MLT. Additionally, the local translation classifier only

Table V
EXAMPLE OF THE TOP-5 MOST SIMILAR SECTIONS FOR 5 LANGUAGES (SPANISH, JAPANESE, ARABIC, RUSSIAN AND FRENCH) TO ENGLISH. IN BOLD THE GROUND-TRUTH.

Language	Query section	Candidates
Spanish	Discografía	**Discography** Recordings Selected discography Albums Solo discography
Japanese	ディスコグラフィ	**Discography** Selected discography Singles Solo discography Albums
Arabic	ديسكوغرافيا	Discography **Albums** Singles Other charted songs Awards
Russian	Дискография	**Discography** Selected discography Recordings Singles Albums
French	Discographie	**Discography** Selected discography Recordings Albums Partial discography

marginally outperformed the global classifier in two language pairs (by 0.02 and 0.01 respectively).

Fig. 3 shows the performance of the global classifier without synonymy detection as we increase the precision threshold from 1 to 3 and 5. In this figure we observe that as the threshold increases, the range of the performance of the global classifier improves from [0.23, 0.73] for precision at 1 to [0.62, 0.96] for precision at 5.

Global translation classifier with synonymy detection (GTC). Fig. 4 shows the performance of the global translation classifier when synonymy detection is included. By comparing the results of Fig. 3 and Fig. 4 it is immediately visible that including the synonymy classifier improves precision at 1, 3, and 5. We specifically see an average improvement of 14% for precision @1, 10% for precision @3, and 6% for precision @5.

In Table V we provide examples of the translations obtained. For convenience, we show the translation from the other five languages to English (to make it easier for the English reader to understand the list of candidates). By doing a qualitative analysis, we found that the top-5 list tends to be semantically consistent, including synonyms (DISCOGRAPHY VS ALBUMS), plural vs singular, and overlapped versions of the same concept (DISCOGRAPHY VS PARTIAL DISCOGRAPHY). This behavior is consistent across all language pairs.

VII. DISCUSSION AND FUTURE WORK

With 45 million articles across 160 language editions, Wikipedia contains a vast body of human knowledge. However, each Wikipedia language edition, independent of its size, contains a significant amount of information not available in other Wikipedia languages [30], [31]. Identifying what content is missing across Wikipedia languages and what content can be relatively easily surfaced to readers of the languages that currently miss such content requires building an alignment across Wikipedia articles that is more granular and precise than article alignment and yet higher level than word by word alignment. In this research we propose aligning sections across Wikipedia languages as the middle ground.

Probabilistic Soft Logic. Using Statistical Relational Learning methods such as PSL always come at the cost of computational performance. The templated rules in PSL leads to graphs that are polynomial in the size of the input data. In addition, the joint nature of its inference is what makes PSL well suited to this task, but it also makes PSL slower than independent and identically distributed (IID) models. Recent work on the performance of SRL systems [32] suggest that blocking, a means of inducing sparsity in the inference graph, is critical to scalability. Our current method of blocking is by only including a translation edge if there exists at least one Wikidata concept c such that s^L appears in language L's version of c and s^M also appears in language M's version of c. This method works well when the Wikidata concept graph is complete and has little noise. However, this blocking scheme becomes less accurate when there are articles that are unique to a specific language, or when a section title is relatively rare within a language. Improving the blocking scheme can help catch translations that are not currently included in the graph, and improve already existing translations through joint inference. Exploring different blocking strategies to use in place of or in conjunction with the current blocking strategy can help to improve the results of the global translation classifier. The potential of multilingual translation using PSL should be considered for future studies in this field.

Synonym classifier. For building the local synonym classifier (Sec. IV-B) we have relied on a state-of-art technique, namely, word-embeddings. Additional features marginally improved the classifier performance. While word-embeddings provide a meaningful metric for semantic distance, determining the actual threshold for classifying two section titles as synonyms is a difficult task. Given that PSL works with numerical input, we decided to use probability of being a synonym as input for our model. Future research can look into developing a reliable standalone synonym classifier.

The trade-off between precision and recall. Although our model obtains a good performance in the majority of the cases, there are a few language pairs (such as Japanese to French or Russian to Japanese) for which the baseline performs better or that the precision@1 is below 0.4. However, in these cases the model performance improves significantly when precision@3 or @5 is considered. As mentioned in

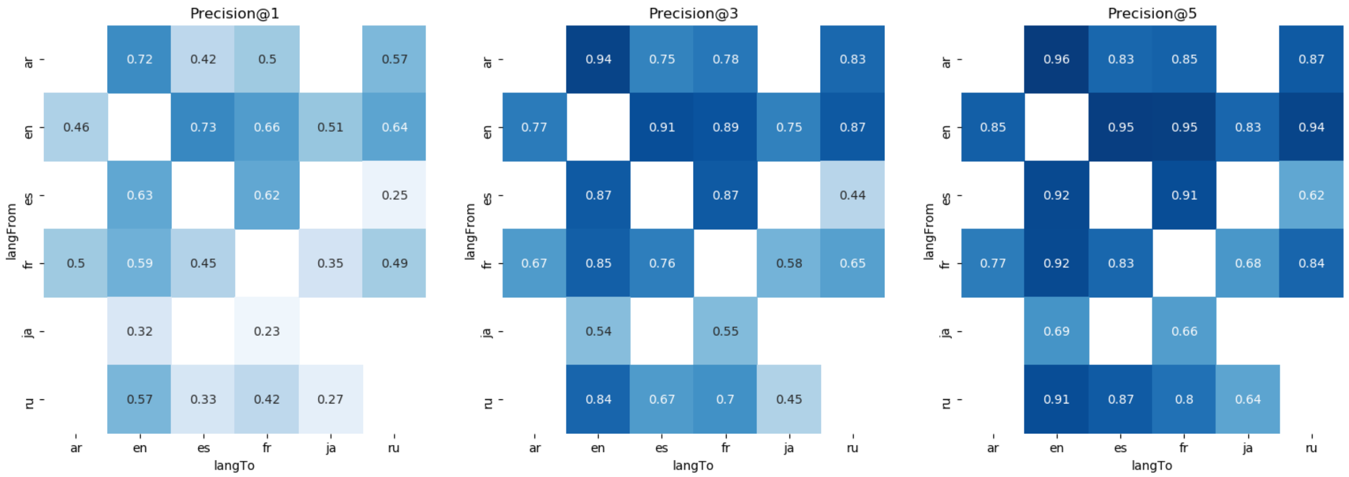


Figure 3. Precision at 1, 3, and 5 for Global Classifier without Synonymy classifier. Empty boxes reflect no ground truth labels for that pair. Note that the leftmost matrix ("Precision@1") corresponds to column "GTC w/o syn" in Table III.

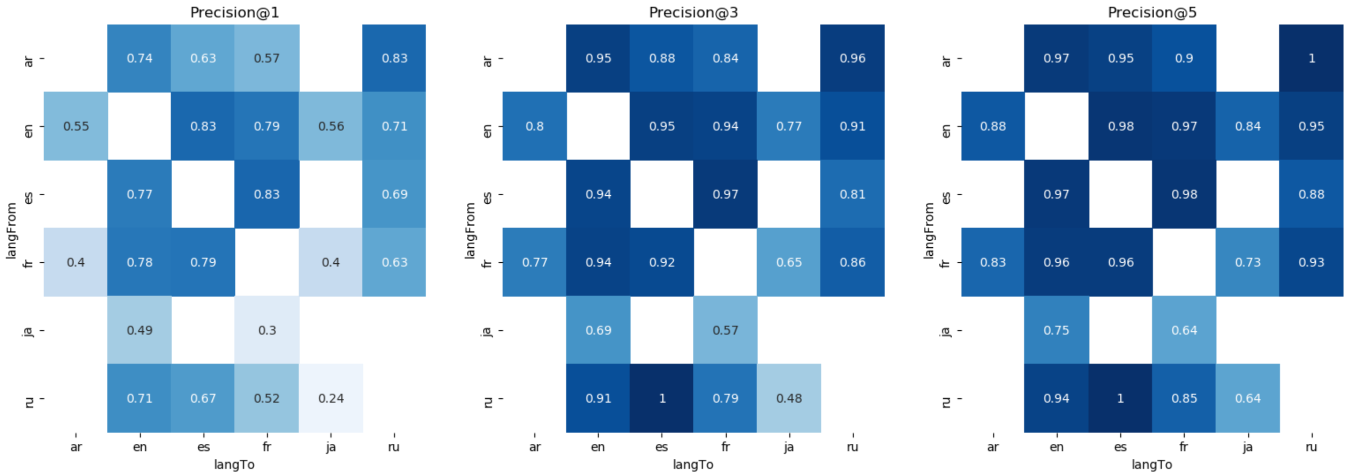


Figure 4. Precision at 1, 3, and 5 for Global Classifier with Synonymy classifier. Empty boxes reflect no ground truth labels for that pair. Note that the leftmost matrix ("Precision@1") corresponds to column "GTC" in Table III.

past research [28], human centered systems designed for Wikipedia can tolerate low precision in exchange for high recall. In real life applications of the model developed in this research, such as a section title recommender system [28], this implies that showing three or five recommendations is sufficient for capturing the exact match. We hypothesize that the majority of the differences between the top-1 and top-5 sections recommended for alignment with a given section are small and the improvement in performance from @1 to @5 is the consequence of the low recall of our ground truth. To prove this hypothesis, a new round of human label collections is required in which all the top-5 sections are evaluated by experienced editors.

Building large-scale multilingual datasets. One of the biggest challenges of this study was to collect synonym and translation labels from *experienced editors* in the six languages of the study and across the 30 language pairs. In some language pairs such as Spanish to Japanese or Russian to Arabic we simply did not find enough editors to help us with the labeling tasks. Automatic collection of section titles as part of the usual workflow of editors when translating through extensions such as Content Translation[10], or large-scale crowd-sourced mappings, e.g., using Amazon Mechanical Turk, can be incorporated in future iterations. It is also important to remark that some of the instances of lower performance for the global translation model are from language pairs that were explicitly chosen for being linguistically distant and difficult to translate between (*i.e.*, Arabic, Japanese, and Russian). In reality, direct translation between these languages in the Wikipedia environment is rare and the strong performance at higher precision tolerances (@3 and @5) along with good performance between the distant languages and at least one other language (usually English) provide various workarounds for language pairs with lower performance.

[10] The Content Translation tool allows editors to create translations to existing articles: https://www.mediawiki.org/wiki/Content_translation

Feature improvement. Future research can investigate improvements of the current model through the addition of new features or considering other distance metrics such as *Cross-domain Similarity Local Scaling (CSLS)* used by [14] for comparing embeddings. To improve the embeddings, a potential extension of this work can also include the use of multi-modal entity representation methods that fuse cross-lingual content, image data, and Knowledge Graphs [33].

Beyond Wikipedia. While the section title alignment problem described in this paper is specific to Wikipedia, the impact of this research goes beyond Wikipedia itself. Future research can leverage aligned *section instances* in multiple ways, for example, to train translation algorithms based on Wikipedia across many languages, or to improve multi-lingual entity linking by pooling recommended article links across languages [34].

VIII. Conclusions

We have defined the *section title* alignment problem and proposed a solution that outperforms a machine translation tool in roughly 80% of the cases. We have modeled Wikipedia sections as language-independent vectors and developed a multi-stage Machine Learning framework that combines crowdsourced labels, Gradient Boosting Trees, and Probabilistic Soft Logic to align section titles across languages. We have shown that a multilingual approach with joint reasoning about multiple translations improves the overall performance of the alignment. To the best of our knowledge, this is the first use of Probabilistic Soft Logic for a multilingual content alignment task. The code, API, and datasets containing labels for translations in six different languages and language-independent vector representations of Wikipedia sections are shared publicly for future related research.

References

[1] W. Statistics, "Wikimedia statistics," 2019. [Online]. Available: https://stats.wikimedia.org/v2/#/all-wikipedia-projects

[2] B. J. Hecht, "The mining and application of diverse cultural perspectives in user-generated content," Ph.D. dissertation, Northwestern University, 2013.

[3] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[4] Y. Lin, B. Yu, A. Hall, and B. J. Hecht, "Problematizing and addressing the article-as-concept assumption in wikipedia." in *CSCW*, 2017, pp. 2052–2067.

[5] S. Gottschalk and E. Demidova, "Multiwiki: interlingual text passage alignment in wikipedia," *ACM Transactions on the Web (TWEB)*, vol. 11, no. 1, p. 6, 2017.

[6] W. Hwang, H. Hajishirzi, M. Ostendorf, and W. Wu, "Aligning sentences from standard wikipedia to simple wikipedia," in *HLT-NAACL*. The Association for Computational Linguistics, 2015, pp. 211–217.

[7] S. H. Bach, B. Huang, J. Boyd-Graber, and L. Getoor, "Paired-dual learning for fast training of latent variable hinge-loss mrfs," in *International Conference on Machine Learning (ICML)*, 2015, stephen Bach and Bert Huang contributed equally.

[8] P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[9] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*. ACL, 2014, pp. 1724–1734.

[10] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *CoRR*, vol. abs/1309.4168, 2013.

[11] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in *EMNLP*. The Association for Computational Linguistics, 2016, pp. 2289–2294.

[12] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation." Association for Computational Linguistics, 2014.

[13] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," in *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track.*, 2015.

[14] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *ICLR (Poster)*. OpenReview.net, 2018.

[15] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *ICLR 2017*, 2017.

[16] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "Hinge-loss markov random fields and probabilistic soft logic," *Journal of Machine Learning Research (JMLR)*, vol. 18, pp. 1–67, 2017. [Online]. Available: https://github.com/stephenbach/bach-jmlr17-code

[17] J. Pujara and L. Getoor, "Generic statistical relational entity resolution in knowledge graphs," in *Sixth International Workshop on Statistical Relational AI*. IJCAI, 2016.

[18] J. Pujara, H. Miao, L. Getoor, and W. W. Cohen, "Knowledge graph identification," in *ISWC (1)*, ser. Lecture Notes in Computer Science, vol. 8218. Springer, 2013, pp. 542–557.

[19] J. Pujara, B. London, and L. Getoor, "Budgeted online collective inference," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, ser. UAI'15. Arlington, Virginia, USA: AUAI Press, 2015, p. 712–721.

[20] J. Foulds, S. Kumar, and L. Getoor, "Latent topic networks: A versatile probabilistic programming framework for topic models," in *International Conference on Machine Learning (ICML)*, 2015.

[21] P. Bao, B. J. Hecht, S. Carton, M. Quaderi, M. S. Horn, and D. Gergle, "Omnipedia: bridging the wikipedia language gap," in *CHI*. ACM, 2012, pp. 1075–1084.

[22] N. Ostapuk, D. Difallah, and P. Cudré-Mauroux, "ParaGraph: Mapping wikidata tail entities to wikipedia paragraphs," in *IEEE BigData*. IEEE, 2022.

[23] G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl, "Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data," in *WSDM*. ACM, 2012, pp. 53–62.

[24] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *PVLDB*, vol. 5, no. 11, pp. 1483–1494, 2012.

[25] G. Demartini, D. Difallah, and P. Cudré-Mauroux, "Large-scale linked data integration using probabilistic reasoning and crowdsourcing," *VLDB J.*, vol. 22, no. 5, pp. 665–687, 2013.

[26] A. Assi, H. Mcheick, A. Karawash, and W. Dhifli, "Context-aware instance matching through graph embedding in lexical semantic space," *Knowl. Based Syst.*, vol. 186, 2019.

[27] M. Chen, Y. Tian, K. Chang, S. Skiena, and C. Zaniolo, "Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment," in *IJCAI*, 2018, pp. 3998–4004.

[28] T. Piccardi, M. Catasta, L. Zia, and R. West, "Structuring wikipedia articles with section recommendations," in *SIGIR*. ACM, 2018, pp. 665–674.

[29] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[30] E. Filatova, "Multilingual wikipedia, summarization, and information trustworthiness," in *SIGIR workshop on information access in a multilingual world*, vol. 3, 2009.

[31] B. J. Hecht and D. Gergle, "The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context," in *CHI*. ACM, 2010, pp. 291–300.

[32] E. Augustine and L. Getoor, "A comparison of bottom-up approaches to grounding for templated markov random fields," in *SysML*, 2018.

[33] M. Luggen, J. Audiffren, D. Difallah, and P. Cudré-Mauroux, "Wiki2prop: A multimodal approach for predicting wikidata properties from wikipedia," in *WWW*. ACM / IW3C2, 2021, pp. 2357–2366.

[34] M. Gerlach, M. Miller, R. Ho, K. Harlan, and D. Difallah, "Multilingual entity linking system for wikipedia with a machine-in-the-loop approach," in *CIKM*. ACM, 2021, pp. 3818–3827.