The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

280 Birds with One Stone: Inducing Multilingual Taxonomies from Wikipedia Using Character-Level Classification

Amit Gupta, Rémi Lebret, Hamza Harkous, Karl Aberer
École Polytechnique Fédérale de Lausanne
Route Cantonale, 1015 Lausanne

Abstract

We propose a novel fully-automated approach towards inducing multilingual taxonomies from Wikipedia. Given an English taxonomy, our approach first leverages the interlanguage links of Wikipedia to automatically construct training datasets for the *is-a* relation in the target language. Character-level classifiers are trained on the constructed datasets, and used in an optimal path discovery framework to induce high-precision, high-coverage taxonomies in other languages. Through experiments, we demonstrate that our approach significantly outperforms the state-of-the-art, heuristics-heavy approaches for six languages. As a consequence of our work, we release presumably the largest and the most accurate multilingual taxonomic resource spanning over 280 languages.

1 Introduction

Motivation. Machine-readable semantic knowledge in the form of taxonomies (i.e., a collection of *is-a* or *hypernym* edges) has proved to be beneficial in an array of Natural Language Processing (NLP) tasks, including inference, textual entailment, question answering, and information extraction (Biemann 2005). This has led to multiple large-scale manual efforts towards taxonomy induction such as WordNet (Miller 1994). However, manual construction of taxonomies is time-intensive, usually requiring massive annotation efforts. Furthermore, the resulting taxonomies suffer from low coverage and are unavailable for specific domains or languages. Therefore, in the recent years, there has been substantial interest in inducing taxonomies automatically, either from unstructured text (Velardi, Faralli, and Navigli 2013), or from semi-structured collaborative content such as Wikipedia (Hovy, Navigli, and Ponzetto 2013).

Wikipedia, the largest publicly-available source of multilingual, semi-structured content (Remy 2002), has served as a key resource for automated knowledge acquisition. One of its core components is the Wikipedia Category Network (hereafter referred to as **WCN**), a semantic network which links Wikipedia entities (also referred to as Wikipedia pages), such as *Johnny Depp*, with inter-connected categories of different granularity (e.g., *American actors*, *Film actors*, *Hollywood*). The semi-structured nature of WCN has enabled the acquisition of large-scale taxonomies using lightweight rule-based approaches (Hovy, Navigli, and Ponzetto 2013), thus leading to a consistent body of research in this direction.

The first line of work on taxonomy induction from Wikipedia mainly focuses on the English language. This includes WikiTaxonomy (Ponzetto and Strube 2008), WikiNet (Nastase et al. 2010), YAGO (Suchanek, Kasneci, and Weikum 2007; Hoffart et al. 2013), DBPedia (Auer et al. 2007), and Heads Taxonomy (Gupta et al. 2016).

The second line of work aims to exploit the multilingual nature of Wikipedia. MENTA (de Melo and Weikum 2010), one of the largest multilingual lexical knowledge bases, is constructed by linking WordNet and Wikipedias of different languages into a single taxonomy. Similarly, YAGO3 (Mahdisoltani, Biega, and Suchanek 2015) extends YAGO by linking Wikipedia entities in multiple languages with WordNet. The most recent approach to multilingual taxonomy induction from Wikipedia is the Multilingual Wikipedia Bitaxonomy Project or MultiWiBi (Flati et al. 2016). MultiWiBi first induces taxonomies for English, which are further projected to other languages using a set of complex heuristics that exploit the interlanguage links of Wikipedia. Unlike MENTA and YAGO3, MultiWiBi is self-contained in Wikipedia, i.e., it does not require labeled training examples or external resources such as WordNet or Wikitionary. While MultiWiBi is shown to outperform MENTA and YAGO3 considerably, it still achieves low precision for non-English pages that do not have an interlanguage link to English (e.g., 59% for Italian).

Contributions. In this paper, we propose a novel approach towards inducing multilingual taxonomies from Wikipedia. Our approach is fully-automated and language-independent. It provides a significant advancement over state of the art in multilingual taxonomy induction from Wikipedia because of the following reasons:

- Most previous approaches such as MENTA or MultiWiBi rely on a set of complex heuristics that utilize custom hand-crafted features. In contrast, our approach employs text classifiers in an optimal path search framework to induce taxonomies from the WCN. The training set for text classifiers is automatically constructed using the Wikipedia interlanguage links. As a result, our approach is simpler, more principled and easily replicable.
- Our approach significantly outperforms the state-of-the-art

approaches across multiple languages in both (1) standard edge-based precision/recall measures and (2) path-quality measures. Furthermore, our taxonomies have significantly higher branching factor than the state-of-the-art taxonomies without incurring any loss of precision.

- As a consequence of our work, we release presumably the largest and the most accurate multilingual taxonomic resource spanning over 280 languages. We also release edge-based gold standards for three different languages (i.e., French, Italian, Spanish) and annotated path datasets for six different languages (i.e., French, Italian, Spanish, Chinese, Hindi, Arabic) for further comparisons and benchmarking purposes. Our source code, taxonomies, and the evaluation datasets are available at `github.com/amitgupta151/MultiTax`.

2 Taxonomy Induction

Background. We start by providing a description of the various components of Wikipedia, which will aid us in presenting the rest of this paper:

- A Wikipedia **page** describes a single entity or a concept. Examples of pages include *Johnny Depp*, *Person*, or *Country*. Currently, Wikipedia consists of more than 44 million pages spanning across more than 280 different languages (Wikipedia 2017).
- A Wikipedia **category** groups related pages and other categories into broader categories. For example, the category *American actors* groups pages for American actors, such as *Johnny Depp*, as well as other categories, such as *American child actors*. The directed graph formed by pages and categories as nodes, and the groupings as edges is known as the **Wikpedia Category Network (WCN)**. A different WCN exists for each of the 280 languages of Wikipedia. WCN edges tend to be noisy, and are usually a mix of *is-a* (e.g., *Johnny Depp*→*American actors*) and *not-is-a* edges (e.g., *Johnny Depp*⇝*Hollywood*).
- An **Interlanguage link** connects a page (or a category) with their equivalent page (or category) across different languages. For example, the English page for *Johnny Depp* is linked to its equivalent versions in 49 different languages such as French, German or Russian. Two nodes linked by an interlanguage link are hereafter referred to as ***equivalent*** to each other.

Algorithm. We now describe our approach for inducing multilingual taxonomies from the WCN. Given (1) a unified taxonomy of pages and categories in English (we use Heads Taxonomy publicly released by Gupta et al. (2016)[1]), (2) the interlanguage links, and (3) a target language, our approach aims to induce a unified taxonomy of pages and categories for the target language. Our approach runs in three phases:

i) **Projection phase**: create a high-precision, low-coverage taxonomy for the target language by projecting *is-a* edges from the given English taxonomy using the interlanguage links.

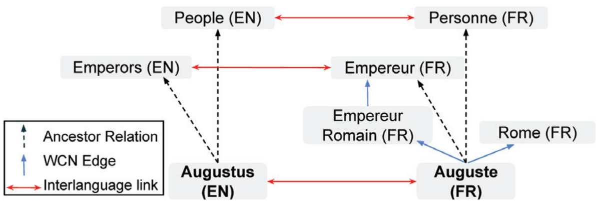


Figure 1: Example of projection phase.

ii) **Training phase**: leverage the high-precision taxonomy to train classifiers for classifying edges into *is-a* or *not-is-a* in the target language.

iii) **Induction Phase**: induce the final high-precision, high-coverage taxonomy by running optimal path search over the target WCN with edge weights computed using the trained classifiers.

2.1 Projection Phase

Let T_e be the given English taxonomy. Let G_f be the WCN and T_f be the output taxonomy (initially empty) for the target language f (such as French). For a node $n_f \in G_f$ with the English equivalent n_e, for which no hypernym exists yet in T_f, we perform the following steps:

i) Collect the set A_e of all ancestor nodes of n_e in T_e up to a fixed height k_1[2].

ii) Fetch the set A_f of equivalents for nodes in A_e in the target language f.

iii) Find the shortest path between n_f and any node in A_f up to a fixed height k_2[3];

iv) Add all the edges in the shortest path to the output taxonomy T_f.

If no English equivalent n_e exists, the node n_f is ignored. Figure 1 shows an example of the projection phase with French as the target language. For French node *Auguste*, its English equivalent (i.e., *Augustus*) is fetched via the interlanguage link. The ancestors of *Augustus* in English taxonomy (i.e., *Emperors*, *People*) are collected, and mapped to their French equivalents (i.e., *Empereur*, *Personne*). Finally, the WCN edges in the shortest path from *Auguste* to *Empereur* (i.e., *Auguste*→*Empereur Romain*, *Empereur Romain*→*Empereur*) are added to the output French taxonomy.

2.2 Training Phase

Up till now, we constructed an initial taxonomy for the target language by simply projecting the English taxonomy using the interlanguage links. However, the resulting taxonomy suffers from low coverage, because nodes that do not have an English equivalent are ignored. For example, only 44.8% of the entities and 40.5% of the categories from the French WCN have a hypernym in the projected taxonomy.

Therefore, to increase coverage, we train two different binary classifiers for classifying remaining target WCN

[1] We note that our method is independent of the English taxonomy induction method.

[2] In our experiments, $k_1 = 14$ sufficed as Heads taxonomy had a maximum height of 14 and no cycles.

[3] k_2 is set to 3 to maintain high precision.

edges into *is-a* (positive) or *not-is-a* (negative). The first classifier is for Entity→Category edges and the other for Category→Category edges[4]. We construct the training data for edge classification as follows:

i) Assign an *is-a* label to the edges in T_f (i.e., the projected target taxonomy).

ii) Assign a *not-is-a* label to all the edges in G_f (i.e., the target WCN) that are not in T_f but originate from a node covered in T_f.

For example, in Figure 1, the edge *Auguste→Empereur Romain* is assigned the *is-a* label, and other WCN edges starting from *Auguste* (e.g., *Auguste→Rome*) are assigned the *not-is-a* label. We note that the *not-is-a* labels, which are assigned during this phase, are not final; they are only temporarily assigned for training the edge classifiers. The final labels are assigned in the next phase (i.e., the induction phase). While, some edges that are assigned temporary *not-is-a* labels may actually be correct *is-a* edges, this design ensures that most of the edges with the assigned *is-a* label, are correct *is-a* edges, thus leading to training of classifiers that achieve high accuracy.

Classifiers. To classify edges into *is-a* or *not-is-a*, we experiment with the following classifiers:

i) **Bag-of-words TFIDF**: Given edge $A \rightarrow B$, concatenate the features vectors for A and B computed using TFIDF over bag of words, and train a linear Support Vector Machine over the concatenated features. This method is hereafter referred to as **Word TFIDF**.

ii) **Bag-of-character-ngrams TFIDF:** Same as Word TFIDF, except TFIDF is computed over bag of character n-grams[5] (hereafter referred to as **Char TFIDF**).

iii) **fastText:** A simple yet efficient baseline for text classification based on a linear model with a rank constraint and a fast loss approximation. Experiments show that fastText typically produces results on par with sophisticated deep learning classifiers (Grave et al. 2017).

iv) **Convolutional Neural Network (CNN):** We use a single-layer CNN model trained on top of word vectors as proposed by Kim (2014). We also experiment with a character version of this model, in which instead of words, vectors are computed using characters and fed into the CNN. These models are referred to as **Word CNN** and **Char CNN** respectively. Finally, we experiment with a two-layer version of the character-level CNN proposed by (Zhang, Zhao, and LeCun 2015), hereafter referred to as **Char CNN-2l**.

v) **Long Short Term Memory Network (LSTM):** We experiment with both word-level and character-level versions of LSTM (Hochreiter and Schmidhuber 1997). These models are hereafter referred to as **Word LSTM** and **Char LSTM** respectively.

2.3 Induction Phase

In the last step of our approach, we discover taxonomic edges for nodes not yet covered in the projected taxonomy (T_f). To this end, we first set the weights of Entity→Category and Category→Category edges in the target WCN as the probability of being *is-a* (computed using the corresponding classifiers). Further, for each node n_f without a hypernym in T_f, we find the top k paths[6] with the highest probabilities originating from n_f to any node in T_f, where the probability of a path is defined as the product of probabilities of individual edges[7]. The individual edges of the most probable paths are added to the final taxonomy.

3 Evaluation

In this section, we compare our approach against the state of the art using two different evaluation methods. In Section 3.1, we compute standard edge-level precision, recall, and coverage measures against a gold standard for three languages. In section 3.2, we perform a comprehensive path-level comparative evaluation across six languages. We compare our approach against MultiWiBi due to the following reasons:

- Only MENTA, MultiWiBi, and our taxonomies are constructed in a fully language-independent fashion; hence, they are available for all 280 Wikipedia languages.
- Unlike YAGO3, MENTA and most other approaches, MultiWiBi and ours are self-contained in Wikipedia. They do not require manually labeled training examples or external resources, such as WordNet or Wikitionary.
- MultiWiBi has been shown to outperform all other previous approaches including YAGO3 and MENTA (Flati et al. 2016).

3.1 Edge-level Evaluation

Experimental Setup. We faced a tough choice of selecting a Wikipedia snapshot since MultiWiBi, to which we compare, is constructed using a 2012 snapshot whereas Gupta et al. (2016), on which we build, uses a 2015 snapshot. Additionally, the code, executable, and gold standards used by MultiWiBi were not available upon request. Therefore, to advance the field and produce a more recent resource, we decided to use a 2015 snapshot of Wikipedia, especially given that Gupta et al. (2016) point out that there is no evidence that taxonomy induction is easier on recent editions of Wikipedia.

We create gold standards for three languages (French, Spanish and Italian) by selecting 200 entities and 200 categories randomly from the November 2015 snapshot of Wikipedia and annotating the correctness of the WCN edges originating from them[8]. Table 1 shows a sample of annotated edges from the French gold standard. In total, 4045 edges were annotated across the three languages.

[4] Entity→Entity and Category→Entity edges are not present in the WCN.

[5] n-values={2,3,4,5,6} worked best in our experiments.

[6] k is set to 1 unless specified otherwise.

[7] If multiple paths with the same probabilities are found, the shortest paths are chosen.

[8] Two annotators independently annotated each edge. Inter-annotator agreement (Cohen's Kappa) varied between 0.71 to 0.93 for different datasets.

is-a
Naissance à Omsk→Naissance en Russie par ville
Port d'Amérique du Sud→Port par continent
not-is-a
Naissance à Omsk⇝Omsk
Port d'Amérique du Sud⇝Géographie de l'Amérique du Sud

Table 1: Examples of Annotated Edges (French).

For evaluation, we use the same metrics as MultiWiBi: (1) Macro-precision (P) defined as the average ratio of correct hypernyms to the total number of hypernyms returned (per node), (2) Recall (R) defined as the ratio of nodes for which at least one correct hypernym is returned, and (3) Coverage (C) defined as the ratio of nodes with at least one hypernym returned irrespective of its correctness.

Training Details. All neural network models are trained on Titan X (Pascal) GPU using the Adam optimizer (Kingma and Ba 2014). Grid search is performed to determine the optimal values of hyper-parameters. For CNN models, we use an embedding of 50 dimensions. The number of filters is set to 1024 for word-level models and 512 for character-level models. For Char CNN-2l model, we use the same parameters used in Zhang, Zhao, and LeCun (2015). For LSTM models, we use an embedding of 128 dimensions, and 512 units in the LSTM cell. We also experimented with more complex architectures, such as stacked LSTM layers and bidirectional LSTMs. However, these architectures failed to provide any significant improvements over the simpler ones.

		Entity			Category		
Language	Method	P	R	C	P	R	C
French	Original WCN	72.0	100	100	78.8	100	100
	MENTA	81.4	48.8	59.8	82.6	55.0	65.7
	MultiWiBi	84.5	80.9	94.1	80.7	80.7	100
	UNIFORM	80.6	83.2	100	85.7	86.7	100
	Word TFIDF	86.5	90.1	100	82.1	83.1	100
	Char TFIDF	**88.0**	**91.7**	100	92.3	93.4	100
	fastText	86.5	90.1	100	90.5	91.6	100
	Word LSTM	**87.8**	**91.5**	100	91.6	92.7	100
	Char LSTM	86.2	89.8	100	**93.9**	**95.1**	100
	Word CNN	86.3	90.0	100	**92.8**	**93.9**	100
	Char CNN	86.2	89.9	100	**93.3**	**94.4**	100
	Char CNN-2l	**87.7**	**91.0**	100	92.2	93.3	100
Italian	Original WCN	74.5	100	100	76.2	100	100
	MENTA	79.7	53.2	66.7	77.1	25.4	32.8
	MultiWiBi	80.1	79.4	96.3	**89.7**	**89.0**	99.2
	UNIFORM	77.7	81.6	100	86.6	88.3	100
	Word TFIDF	**90.0**	**94.4**	100	84.1	85.7	100
	Char TFIDF	88.4	92.8	100	**89.2**	**90.9**	100
	fastText	86.8	91.1	100	**87.3**	**89.0**	100
	Word LSTM	**90.9**	**95.4**	100	83.1	84.8	100
	Char LSTM	89.8	94.4	100	83.3	83.8	100
	Word CNN	89.6	94.3	100	83.1	84.8	100
	Char CNN	**92.6**	**97.2**	100	86.9	88.7	100
	Char CNN-2l	87.7	92.1	100	86.1	87.8	100
Spanish	Original WCN	81.4	100	100	80.9	100	100
	MENTA	81.0	42.9	52.7	80.5	54.2	66.4
	MultiWiBi	87.0	82.0	93.7	84.8	84.4	100
	UNIFORM	88.0	90.7	100	83.0	85.0	100
	Word TFIDF	89.9	92.7	100	78.9	80.8	100
	Char TFIDF	92.5	95.4	100	88.3	90.4	100
	fastText	**93.0**	**95.9**	100	**88.9**	**91.0**	100
	Word LSTM	**93.4**	**96.3**	100	88.2	90.3	100
	Char LSTM	92.3	95.3	100	88.8	90.3	100
	Word CNN	92.9	95.8	100	87.6	89.7	100
	Char CNN	92.9	95.8	100	**92.9**	**95.1**	100
	Char CNN-2l	**93.3**	**96.3**	100	**89.9**	**92.1**	100

Table 2: Edge-level precision (P), recall (R) and Coverage (C) scores for different methods. MENTA and MultiWiBi results as reported by Flati et al. (2016). The top 3 results are shown in bold, and the best is also underlined.

Results. Table 2 shows the results for different methods including the state-of-the-art approaches (i.e., MENTA and MultiWiBi) and multiple versions of our three-phase approach with different classifiers. It also includes two baselines, i.e., **WCN** and **UNIFORM**. The WCN baseline outputs the original WCN as the induced taxonomy without performing any kind of filtering of edges. UNIFORM is a uniformly-random baseline, in which all the edge weights are set to 1 in the induction phase (cf. Section 2.3).

Table 2 shows that all classifiers-based models achieve significantly higher precision than UNIFORM and WCN baselines, thus showing the utility of weighing with classification probabilities in the Induction phase. Interestingly, UNIFORM achieves significantly higher precision than WCN for both entities and categories across all three languages, hence, demonstrating that optimal path search in the Induction phase also contributes towards hypernym selection. All classifier-based approaches (except Word TFIDF) significantly outperform MultiWiBi for entities across all languages as well as for French and Spanish categories. Although MultiWiBi performs better for Italian categories, Char TFIDF achieves similar performance (89.2% vs 89.7%) [9].

Coverage is 100% for all the baselines and the classifiers-based approaches. This is because at least one path is discovered for each node in the induction phase, thus resulting in at least one (possibly incorrect) hypernym for each node in the final taxonomy. This also serves to demonstrate that the initial projected taxonomy (cf. Section 2.1) is reachable from every node in the target WCN.

In general, character-level models outperform their word-level counterparts. Char TFIDF significantly outperforms Word TFIDF for both entities and categories across all languages. Similarly, Char CNN outperforms Word CNN. Char LSTM outperforms Word LSTM for categories, but performs slightly worse for entities. We hypothesize that this is due to

[9] We note that entity edges are qualitatively different for MultiWiBi and other methods, i.e., MultiWiBi has Entity→Entity edges whereas other methods have Entity→Category edges. Given that fact and the unavailability of the gold standards from MultiWiBi, we further support the efficacy of our approach with a direct path-level comparison in the next section.

MultiWiBi
Patrimoine mondial en Équateur ⇝ Conservation de la nature → Écologie → Biologie → Sciences naturelles → Subdivisions par discipline → Sciences → Discipline académique → Académie → Concept philosophique

Char TFIDF
Patrimoine mondial en Équateur → **Patrimoine mondial en Amérique** → **Patrimoine mondial par continent** → **Patrimoine mondial** → **Infrastructure touristique** → **Lieu** ⇝ Géographie → Discipline des sciences humaines et sociales → Sciences humaines et sociales → Subdivisions par discipline

Table 3: Samples of generalization paths for French categories from MultiWiBi and Char TFIDF taxonomies. Correct path prefix (CPP) for each path is shown in bold.

the difficulty in training character LSTM models over larger training sets. Entity training sets are much larger, as the number of Entity→Category edges are significantly higher than the number of Category→Category edges (usually by a factor of 10).

Neural Models vs. TFIDF. CNN-based models perform slightly better on average, followed closely by LSTM and TFIDF respectively. However, the training time for neural networks-based models is significantly higher than TFIDF models. For example, it takes approximately 25 hours to train the Char CNN model for French entities using a dedicated GPU. In contrast, the Char TFIDF model for the same data is trained in less than 5 minutes. Therefore, for the sake of efficiency, as well as to ensure simplicity and reproducibility across all languages, we choose Char TFIDF taxonomies as our final taxonomies for the rest of the evaluations. However, it is important to note that more accurate taxonomies can be induced by using our approach with neural-based models, especially if the accuracy of taxonomies is critical for the application at hand.

3.2 Path-level Evaluation

In this section, we compare Char TFIDF against MultiWiBi using a variety of path-quality measures. Path-based evaluation of taxonomies was proposed by Gupta et al. (2016), who demonstrated that good edge-level precision may not directly translate to good path-level precision for taxonomies. They proposed the average length of *correct path prefix (CPP)*, i.e., the maximal correct prefix of a generalization path, as an alternative measure of quality of a taxonomy. Intuitively, it aims to capture the average number of upward generalization hops that can be taken until the first wrong hypernym is encountered. Following this metric, we randomly sample paths originating from 25 entities and 25 categories from the taxonomies, and annotate the first wrong hypernym in the upward direction. In total, we annotated 600 paths across six different languages for Char TFIDF and MultiWiBi taxonomies. Table 3 shows examples of these generalization paths along with their CPPs[10].

		Entity			Category		
Language	Method	AL	ACPP	ARCPP	AL	ACPP	ARCPP
French	MultiWiBi	8.24	2.96	0.49	8.92	3.6	**0.56**
	Char TFIDF	11.08	**5.08**	0.49	8.36	**3.76**	0.49
Italian	MultiWiBi	7.36	2.68	0.45	14.84	3.72	0.27
	Char TFIDF	8.32	**4.88**	**0.61**	8.32	**4.52**	**0.57**
Spanish	MultiWiBi	7.04	3.08	**0.55**	12.08	4.08	0.36
	Char TFIDF	12.8	**5.0**	0.48	12.76	**5.28**	**0.48**
Arabic	MultiWiBi	8.96	2.12	0.31	14.64	4.12	0.31
	Char TFIDF	7.48	**5.88**	**0.81**	6.96	**5.04**	**0.74**
Hindi	MultiWiBi	7.72	1.88	0.27	7.4	1.8	0.36
	Char TFIDF	10.28	**4.92**	**0.47**	8.0	**2.44**	**0.38**
Chinese	MultiWiBi	7.4	2.56	0.47	8.0	4.43	0.63
	Char TFIDF	6.32	**3.92**	**0.68**	6.95	**4.48**	**0.68**

Table 4: Comparison of average path length (AL), average length of correct path prefix (ACPP), and average ratio of CPP to path lengths (ARCPP).

We report the average length of CPP (ACPP), as well as the average ratio of length of CPP to the full path (ARCPP). As an example, given the generalization path ***apple→fruit***⇝*farmer→human→animal* with the *not-is-a* edge *fruit⇝farmer*, the path length is 5, length of CPP is 2, and ratio of length of CPP to total path is 0.4 (i.e., $\frac{2}{5}$).

Table 4 shows the comparative results. Char TFIDF taxonomies significantly outperform MultiWiBi taxonomies, achieving higher average CPP lengths (ACPP) as well as higher average ratio of CPP to path lengths (ARCPP). Therefore, compared to the state-of-the-art MultiWiBi taxonomies, Char TFIDF taxonomies are a significantly better source of generalization paths for both entities and categories across multiple languages.

4 Analysis

In this section, we perform additional analyses to gain further insights into our approach. More specifically, in Section 4.1 and 4.2, we perform an in-depth comparison of the Word TFIDF and Char TFIDF models. In section 4.3, we show the effect of the parameter k, i.e., the number of paths discovered during optimal path search (cf. Induction Phase in Section 2.3), on the branching factor and the precision of the induced taxonomies.

4.1 Word vs. Character Models

To compare word and character-level models, we first report the validation accuracies for Word TFIDF and Char TFIDF models in Figure 2, as obtained during the training phase[11] (cf. Section 2.2). Char TFIDF models significantly outperform Word TFIDF models, achieving higher validation accuracies across six different languages. The improvements are usually higher for languages with non-Latin scripts. This can be partly attributed to the error-prone nature of whitespace-based tokenization for such languages. For example, the word

[10] Same starting entities and categories are used for all taxonomies per language.

[11] Validation set is constructed by randomly selecting 25% of the edges with each label (i.e., *is-a* and *not-is-a*) as discovered during the projection phase.

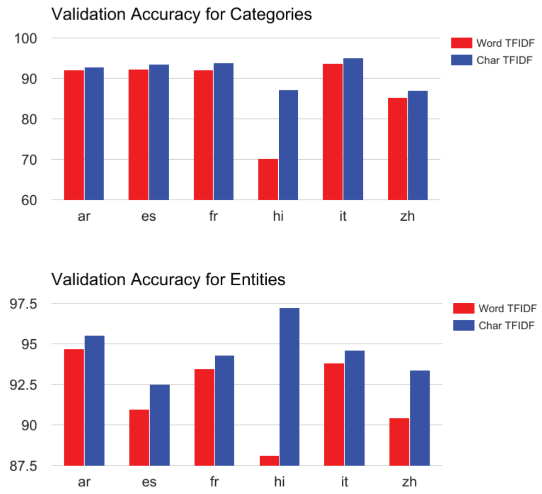
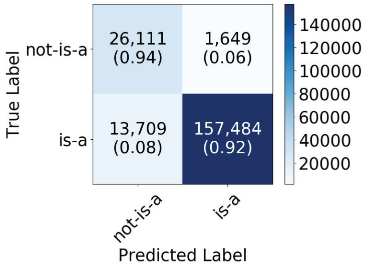
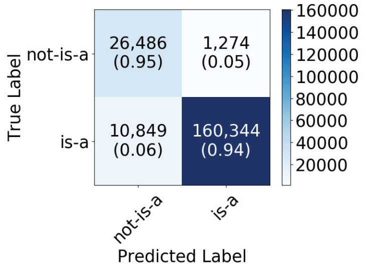


Figure 2: Validation accuracies for Word TFIDF vs. Char TFIDF models.



(a) **Word TFIDF**



(b) **Char TFIDF**

Figure 3: Confusion matrices for Word TFIDF vs. Char TFIDF for French categories. Each cell shows the total number of edges along with the ratios in brackets.

Word TFIDF	Char TFIDF
dolphins, dolphins, miami miami, entraîneur, des	s dol, s dolp, es dol hins, dolph, hins d

Table 5: Top features for *not-is-a* edge *Entraîneur des Dolphins de Miami⇝Dolphins de Miami*.

Word TFIDF	Char TFIDF
dolphins, américain, miami entraîneur, sportif, entraîneur	ur spo, r spor, eur sp tif am, if am, if amé

Table 6: Top features for *is-a* edge *Entraîneur des Dolphins de Miami→Entraîneur sportif américain*.

tokenizer for Hindi splits words at many accented characters in addition to word boundaries, thus leading to erroneous features and poor performance. In contrast, character-level models are better equipped to handle languages with arbitrary scripts, because they do not need to perform text tokenization.

4.2 False Positives vs. False Negatives

To further compare word and character models, we focus on the specific case of French categories. In Figure 3, we show the confusion matrices of Word TFIDF and Char TFIDF model computed using the validation set for French categories. While, in general, both models perform well, Char TFIDF outperforms Word TFIDF, producing fewer false positives as well as false negatives. We noticed similar patterns across most languages for both entities and categories.

We hypothesize that the superior performance of Char TFIDF is because character n-gram features incorporate the morphological properties computed at the sub-word level as well as word boundaries, which are ignored by the word-based features. To demonstrate this, we show in Tables 5 and 6 the top Word TFIDF and Char TFIDF features of a *not-is-a* and an *is-a* edge. These edges are misclassified by Word TFIDF, but correctly classified by Char TFIDF.

While Word TFIDF features are restricted to individual words, Char TFIDF features can capture patterns across word boundaries. For example the 6-gram feature "*ur spo*" occurs in multiple hypernyms with different words: e.g., *Commentateur sportif américain*, *Entraîneur sportif américain* and *Entraîneur sportif russe*. Such features incorporate morphological information such as plurality and affixes, which can be important for the detection of an *is-a* relationship. This is also evidenced by previous approaches that utilize multiple hand-crafted features based on such morphological information (Suchanek, Kasneci, and Weikum 2007; Gupta et al. 2016). Therefore, character-level models equipped with such features perform better at the task of WCN edge classification than their word-level counterparts.

4.3 Precision vs. Branching Factor

Along with standard precision/recall measures, structural evaluation also plays an important role in assessing the quality of a taxonomy. One of the important structural properties of a taxonomy is the *branching factor*, which is defined as the average out-degree of the nodes in the taxonomy. Taxonomies with higher branching factors are desirable, because they are better equipped to account for multiple facets of a concept or an entity (e.g., *Bill Gates* is both a philanthropist and an entrepreneur).

However, there is usually a trade-off between branching factor and precision in automatically induced taxonomies (Velardi, Faralli, and Navigli 2013). Higher branching factor typically results in lowering of precision due to erroneous edges with lower scores being added to the taxonomy. Prioritizing the precision over the branching factor or vice-versa is usu-

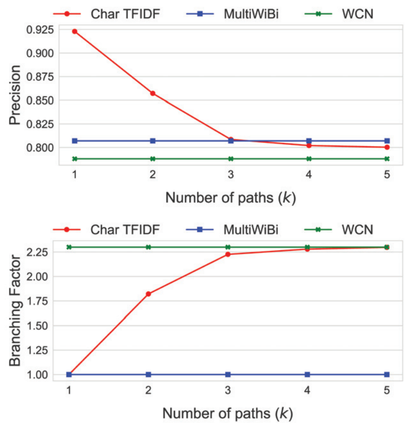


Figure 4: Precision vs. branching factor for different number of paths (k) in the Induction phase (cf. Section 2.3).

ally determined by the specific use case at hand. Therefore, it is desirable for a taxonomy induction method to provide a control mechanism over this trade-off.

In our approach, the number of paths discovered (k) in the optimal path search (cf. Section 2.3) serves as the parameter for controlling this trade-off. As k increases, the branching factor of the induced taxonomy increases because more paths per term are discovered. To demonstrate this effect, we plot the values of precision and branching factor of Char TFIDF taxonomies for varying values of k for French categories[12] in Figure 4. Precision and branching factors for MultiWiBi taxonomies and the original WCN are also shown for comparison purposes.

Char TFIDF significantly outperforms MultiWiBi, either achieving higher precision ($k\leq2$) or higher branching factor ($k\geq2$). At k=2, Char TFIDF presents a sweet spot, outperforming MultiWiBi in both precision and branching factor. For $k\geq3$, Char TFIDF taxonomies start to resemble the original WCN, because most of the WCN edges are selected by optimal path discovery. This experiment demonstrates that in contrast to MultiWiBi's fixed set of heuristics, our approach provides a better control over the branching factor of the induced taxonomies.

5 Related Work and Discussion

The large-scale and high quality of Wikipedia content has enabled multiple approaches towards knowledge acquisition and taxonomy induction over the past decade. Earlier attempts at taxonomy induction from Wikipedia focus on the English language. WikiTaxonomy, one of the first attempts to taxonomize Wikipedia, labels English WCN edges as *is-a* or *not-is-a* using a cascade of heuristics based on hand-crafted features (Ponzetto and Strube 2008). WikiNet extends WikiTaxonomy by expanding *not-is-a* relations into more fine-grained relations such as meronymy (i.e., *part-of*) and geo-location (i.e., *located-in*). YAGO induces a taxonomy by linking Wikipedia categories to WordNet synsets using a set of simple heuristics (Suchanek, Kasneci, and Weikum 2007; Hoffart et al. 2013). DBPedia provides a fully-structured knowledge representation for the semi-structured content of Wikipedia, which is further linked to existing knowledge bases such as YAGO and OpenCyc (Auer et al. 2007; Lehmann et al. 2015). More recently, Gupta et al. (2016) induce a unified taxonomy of entities and categories from English WCN using a novel set of high-precision heuristics that classify WCN edges into *is-a* and *not-is-a*.

A second line of work aims to extend the taxonomy induction process to other languages by exploiting the multilingual nature of Wikipedia content. MENTA, a large-scale multilingual knowledge base, is induced by linking WordNet with WCN of different languages into a unified taxonomy (de Melo and Weikum 2010). The most recent and the most notable effort towards this direction is MultiWiBi (Flati et al. 2016). MultiWiBi first simultaneously induces two separate taxonomies for English, one for pages and one for categories. To this end, it exploits the idea that information contained in pages are useful for taxonomy induction over categories and vice-versa. To induce taxonomies for other languages, MultiWiBi employs a set of complex heuristics, which utilize hand-crafted features (such as textual and network topology features) and a probabilistic translation table constructed using the interlanguage links.

Our approach borrows inspiration from many of the aforementioned approaches. First, similar to WikiTaxonomy and Gupta et al. (2016), our approach also classifies WCN edges into *is-a* or *not-is-a*. Second, similar to MultiWiBi, our approach also projects an English taxonomy into other languages using the interlanguage links. However, unlike these approaches, our approach does not employ any heuristics or hand-crafted features. Instead, it uses text classifiers trained on an automatically constructed dataset to assign edge weights to WCN edges. Taxonomic edges are discovered by running optimal path search over the WCN in a fully-automated and language-independent fashion.

Our experiments show that taxonomies derived using our approach significantly outperform the state-of-the-art taxonomies, derived by MultiWiBi using more complex heuristics. We hypothesize that it is because our model primarily uses categories as hypernyms, whereas MultiWiBi first discovers hypernym lemmas for entities using potentially noisy textual features derived from unstructured text. Categories have redundant patterns, which can be effectively exploited using simpler models. This has also been shown by Gupta et al. (2016), who use simple high-precision heuristics based on the lexical head of categories to achieve significant improvements over MultiWiBi for English.

Additionally, for taxonomy induction in other languages, MultiWiBi uses a probabilistic translation table, which is likely to introduce further noise. The high-precision heuristics of Gupta et al. (2016) are not easily extensible to languages other than English, due to the requirement of a syntactic parser for lexical head detection. In contrast, our approach learns such features from automatically generated training data, hence resulting in high-precision, high-coverage taxonomies for all Wikipedia languages.

Our taxonomies contain more than 1 million *is-a* edges

[12] Similar effects are observed for both entities and categories for all languages.

for 10 languages, and more than 100,000 *is-a* edges for 46 languages. For rest of the languages, taxonomies are smaller (i.e., less than 50,000 *is-a* edges), mainly due to the smaller sizes of their corresponding WCNs. Nonetheless, our approach is still effective as it achieves 100% coverage over the WCNs by design. Our taxonomies are available at github.com/amitgupta151/MultiTax.

6 Conclusion

In this paper, we presented a novel approach towards multilingual taxonomy induction from Wikipedia. Unlike previous approaches which are complex and heuristic-heavy, our approach is simpler, principled and easy to replicate. Taxonomies induced using our approach outperform the state of the art on both edge-level and path-level metrics across multiple languages. Our approach also provides a parameter for controlling the trade-off between precision and branching factor of the induced taxonomies. A key outcome of this work is the release of our taxonomies across 280 languages, which are significantly more accurate than the state of the art and provide higher coverage.

Acknowledgements

This work is supported by a Sinergia Grant by the Swiss National Science Foundation (SNF 147609).

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. G. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, 722–735.

Biemann, C. 2005. Ontology learning from text: A survey of methods. *LDV Forum* 20(2):75–93.

de Melo, G., and Weikum, G. 2010. MENTA: inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, 1099–1108.

Flati, T.; Vannella, D.; Pasini, T.; and Navigli, R. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artif. Intell.* 241:66–102.

Grave, E.; Mikolov, T.; Joulin, A.; and Bojanowski, P. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, 427–431.

Gupta, A.; Piccinno, F.; Kozhevnikov, M.; Pasca, M.; and Pighin, D. 2016. Revisiting taxonomy induction over wikipedia. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, 2300–2309.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Hoffart, J.; Suchanek, F. M.; Berberich, K.; and Weikum, G. 2013. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.* 194:28–61.

Hovy, E. H.; Navigli, R.; and Ponzetto, S. P. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artif. Intell.* 194:2–27.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, 1746–1751.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; and Bizer, C. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6(2):167–195.

Mahdisoltani, F.; Biega, J.; and Suchanek, F. M. 2015. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*.

Miller, G. A. 1994. WORDNET: A lexical database for english. In *Human Language Technology, Proceedings of a Workshop held at Plainsboro, New Jerey, USA, March 8-11, 1994*.

Nastase, V.; Strube, M.; Boerschinger, B.; Zirn, C.; and Elghafari, A. 2010. Wikinet: A very large scale multi-lingual concept network. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Ponzetto, S. P., and Strube, M. 2008. Wikitaxonomy: A large scale knowledge resource. In *ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, Proceedings*, 751–752.

Remy, M. 2002. Wikipedia: The free encyclopedia. *Online Information Review* 26(6):434.

Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, 697–706.

Velardi, P.; Faralli, S.; and Navigli, R. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39(3):665–707.

Wikipedia. 2017. List of wikipedias — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias\&oldid773693902. [Online; accessed 9-April-2017].

Zhang, X.; Zhao, J. J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 649–657.