280 Birds with One Stone: Inducing Multilingual Taxonomies from Wikipedia using Character-level Classification

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

Abstract

We propose a simple, yet effective, approach towards inducing multilingual taxonomies from Wikipedia. Given an English taxonomy, our approach leverages the interlanguage links of Wikipedia followed by character-level classifiers 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

Machine-readable semantic knowledge in the form of taxonomies (i.e., a collection of is-a 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, 1995). However, manual construction of taxonomies is time-intensive, usually requiring huge 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 a substantial interest in inducing taxonomies automatically, either from unstructured text (Velardi et al., 2013), or from semi-structured collaborative content such as Wikipedia (Hovy et al., 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 (or WCN), a semantic network which links Wikipedia entities, 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 et al., 2013), thus leading to a consistent body of research in this direction. The first line of work on taxonomy induction from Wikipedia is mainly focused on English. This includes WikiTaxonomy (Ponzetto and Strube, 2008), WikiNet (Nastase et al., 2010), YAGO (Suchanek et al., 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 et al., 2014) 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). 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 Wiktionary. MultiWiBi first simultaneously induces two separate taxonomies for English, one

---

1In this work, we use the terms is-a and hypernym interchangeably.

2We use Wikipedia page and entity interchangeably.
for pages and one for categories. It builds on the idea that information contained in pages are useful for taxonomy induction over categories and vice-versa. To induce taxonomies for other languages, complex heuristics are employed, which utilize textual features, network topology features, and a probabilistic translation table constructed using the interlanguage links. 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 work, we show that we can achieve a superior performance to the state of the art in multilingual taxonomies with a simple, yet effective model. Similar to MultiWiBi, we start from an English taxonomy. However, instead of relying on a complex set of heuristics for transferring this taxonomy to other languages, we develop a novel language-independent approach that uses character n-grams classifiers. We show that our approach significantly outperforms the state-of-the-art approaches in (1) standard edge-based precision/recall measures over multiple languages and in (2) path-quality measures. In addition to releasing our taxonomies for all 280 Wikipedia languages, we also release the evaluation gold standards for future benchmarking and comparisons.

## 2 Taxonomy Induction

**Background.** We first provide a description of the various components of Wikipedia. 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 across 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 Wikipedia 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).

Finally, interlanguage links connect pages (or categories) with their equivalent pages (or categories) across different languages. For example, the English page for Johnny Depp is linked to its equivalent versions in 49 different languages including French (Johnny Depp) and Russian (Депп, Джонни). Two nodes linked by an interlanguage link are hereafter referred to as equivalent to each other.

**Algorithm.** Given (1) a unified taxonomy of pages and categories in English (we use Heads Taxonomy released by Gupta et al. (2016)\(^3\)), (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 steps:

1. **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.
2. **Training phase:** leverage the high-precision taxonomy to train classifiers for classifying edges into is-a or not-is-a in the target language.
3. **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 classifiers.

**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 \( f \). 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:

(a) Collect the set \( A_e \) of all ancestor nodes of \( n_e \) in \( T_e \) up to a fixed height \( k_1 \)\(^4\).

(b) Fetch the set \( A_f \) of equivalents for nodes in \( A_e \) in the target language \( f \).

(c) Find the shortest path between \( n_f \) and any node in \( A_f \) up to a fixed height \( k_2 \)\(^5\).

(d) Add all the edges in the shortest path to the output taxonomy \( T_f \).

**Algorithm.** Given (1) a unified taxonomy of pages and categories in English (we use Heads Taxonomy released by Gupta et al. (2016)\(^3\)), (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 steps:

1. **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.
2. **Training phase:** leverage the high-precision taxonomy to train classifiers for classifying edges into is-a or not-is-a in the target language.
3. **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 classifiers.

**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 \( f \). 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:

(a) Collect the set \( A_e \) of all ancestor nodes of \( n_e \) in \( T_e \) up to a fixed height \( k_1 \)\(^4\).

(b) Fetch the set \( A_f \) of equivalents for nodes in \( A_e \) in the target language \( f \).

(c) Find the shortest path between \( n_f \) and any node in \( A_f \) up to a fixed height \( k_2 \)\(^5\).

(d) Add all the edges in the shortest path to the output taxonomy \( T_f \).

**Algorithm.** Given (1) a unified taxonomy of pages and categories in English (we use Heads Taxonomy released by Gupta et al. (2016)\(^3\)), (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 steps:

1. **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.
2. **Training phase:** leverage the high-precision taxonomy to train classifiers for classifying edges into is-a or not-is-a in the target language.
3. **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 classifiers.

**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 \( f \). 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:

(a) Collect the set \( A_e \) of all ancestor nodes of \( n_e \) in \( T_e \) up to a fixed height \( k_1 \)\(^4\).

(b) Fetch the set \( A_f \) of equivalents for nodes in \( A_e \) in the target language \( f \).

(c) Find the shortest path between \( n_f \) and any node in \( A_f \) up to a fixed height \( k_2 \)\(^5\).

(d) Add all the edges in the shortest path to the output taxonomy \( T_f \).

Figure 1 shows an example of the projection phase with French as the target language.

---

\(^3\)We note though that our method is independent of the English taxonomy induction method.

\(^4\)In our experiments, \(k_1 = 14\) sufficed as Heads taxonomy had a maximum height of 14 and no cycles.

\(^5\)\(k_2\) is set to 3 to maintain high precision.
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.

**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 usually suffers from low coverage. For example, hypernyms are found for only 44.8% of the entities and 40.5% of the categories in the French WCN.

To increase coverage, we train two different binary classifiers for classifying remaining target WCN 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. We construct the training data for edge classification as follows:

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

(b) 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 edges starting from Auguste (e.g., Auguste→Rome) are assigned the not-is-a label.

For the classifiers, we considered first a linear Support Vector Machine trained over edge features. For each edge $A \rightarrow B$, we experiment with two sets of features:

(a) **Bag-of-words TFIDF**: concatenate the features vectors for $A$ and $B$ computed using TFIDF over bag of words (referred to henceforth as **Word TFIDF**).

(b) **Bag-of-character-ngrams**: concatenate the features vectors for $A$ and $B$ computed using TFIDF over bag of character $n$-grams (referred to as **Char TFIDF**).\(^7\)

**Induction Phase.** In the last step, we discover taxonomic edges for nodes not yet covered in the projected taxonomy ($T_f$). To this end, we first set the weight of Entity→Category and Category→Category edges in the target WCN as the probability of being is-a (computed using the corresponding classifier). Further, for each node $n_f$ without a hypernym in $T_f$, we find the path with highest probability 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. 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 Wiktionary.

- MultiWiBi has been shown to outperform all other previous approaches including YAGO3 and MENTA (Flati et al., 2016).

\(^7\) $n$-values $\{2,3,4,5,6\}$ worked best in our experiments.

\(^8\) If multiple paths with the same probability are found, the shortest path is chosen.
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, or 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 2015 WCN and annotating their correctness. Table 1 shows a sample of annotated edges from the French gold standard. 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) as the ratio of nodes for which at least one correct hypernym is returned, and (3) Coverage (C) as the ratio of nodes with at least one hypernym returned irrespective of its correctness.

Results. Table 2 shows the results for different methods including the original WCN (WCN) and a random baseline (RANDOM) where all probabilities are set to 1 in the induction phase (cf. Section 2). Char TFIDF outperforms Word TFIDF as well as MultiWiBi significantly except for Italian where it achieves similar results. Word TFIDF and Char TFIDF models achieve higher precision than RANDOM, which in turns achieves higher precision than original WCN. This suggests that hypernym selection occurs due to two factors in the induction phase (cf. Section 2), i.e., the classification probabilities and the optimal path search.

Table 1 – Examples of Annotated Edges (French).

| Language | Method   | Entity P | R   | C | Category P | R   | C   |
|----------|----------|----------|-----|---|------------|-----|-----|
| French   | WCN      | 72.0     | 100 | 100 | 78.8       | 100 | 100 |
|          | MultiWiBi| 84.5     | 80.9 | 94.1 | 80.7       | 80.7 | 100 |
|          | RANDOM   | 80.6     | 83.2 | 100 | 85.7       | 86.7 | 97.9 |
|          | Word TFIDF| 86.5    | 90.1 | 100 | 82.1       | 83.1 | 97.9 |
|          | Char TFIDF| 88.0    | 91.7 | 100 | 92.3       | 93.4 | 97.9 |
| Italian  | WCN      | 74.5     | 100 | 100 | 76.2       | 100 | 100 |
|          | MultiWiBi| 80.1     | 79.4 | 96.3 | 89.7       | 89   | 99.2 |
|          | RANDOM   | 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 |
| Spanish  | WCN      | 81.4     | 100 | 100 | 80.9       | 100 | 100 |
|          | MultiWiBi| 87.0     | 82.0 | 93.7 | 84.8       | 84.4 | 100 |
|          | RANDOM   | 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 |

We also note that we have also tried more advanced models, based on neural networks. In particular, we experimented with word-level and character-level classifiers based on convolutional neural networks (similar to those developed by Kim (2014) and Zhang et al. (2015)). We consistently observed that such complex models either performed worse or on par with Char TFIDF in the various languages. Hence, for the sake of favoring the simpler models, we discuss the TFIDF models in the rest of this paper.

3.2 Path-level Evaluation

Gupta et al. (2016) demonstrated that high edgewise precision may not translate to high 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
Table 3 – Samples of generalization paths for French categories. Correct path prefix (CPP) for each path is shown in bold.

| Language | Method | Entity | ACPP | ARCPP | Category | 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 | 7.77  | 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 | 6.52  | 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).

4 Analysis

In this section, we perform additional analyses to gain further insights into our approach. In Section 4.1, we perform an in-depth comparison of the Word TFIDF and Char TFIDF models. In Section 4.2, we explain the differences between MultiWiBi and our approach.

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 (cf. Section 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 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.

False Positives vs. False Negatives. To further compare word and character models, we focus on the specific case of French categories. We show the confusion matrices of Word TFIDF and Char TFIDF model computed using a validation set for French categories in Figure 3.

While, in general, both models are performing well, Char TFIDF consistently outperforms Word TFIDF and Char TFIDF models. In Section 4.2, we explain the differences between MultiWiBi and our approach.
4.2 Comparison with MultiWiBi

In Section 3, we showed that taxonomies derived using our simple TFIDF-based models 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 while MultiWiBi first discovers hypernym lemmas for entities using potentially noisy textual features derived from unstructured text. In fact, categories have redundant patterns, which can be effectively exploited using simpler models. This has also been shown by Gupta et al. (2016), who used simple features based on the lexical head of categories to achieve significant improvements over MultiWiBi for English.

Secondly, due to the use of a probabilistic translation table, MultiWiBi is likely to add further noise for other languages. Since the approach of Gupta et al. (2016) is not easily extensible to other languages due to the requirement of a syntactic parser for lexical head detection, we propose to learn such features from automatically generated training data, hence resulting in high-precision, high-coverage taxonomies for multiple Wikipedia languages.

Overall, given the results, we can safely conclude that our approach produces taxonomies, which are a better source for generalizations for entities and categories across the various Wikipedia languages.

5 Conclusion

In this paper, we have shown that a lightweight classification approach can outperform the state-of-the-art, heuristic-heavy techniques for inducing multilingual taxonomies from Wikipedia. Using both edge-level and path-level metrics, we demonstrate the efficacy of our approach across multiple languages. We also demonstrated that character-based TFIDF models outperform word-based TFIDF models for this task. A key outcome of this work is the release of our taxonomies across 280 languages, which are more accurate than the state of the art and provide higher coverage.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. The semantic web pages 722–735.
Chris Biemann. 2005. Ontology learning from text: A survey of methods. In LDV forum, volume 20, pages 75–93.

Gerard de Melo and Gerhard Weikum. 2010. Menta: Inducing multilingual taxonomies from wikipedia. In Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, pages 1099–1108.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. Artificial Intelligence 241:66–102.

Amit Gupta, Francesco Piccinno, Mikhail Kozhevenkov, Marius Pasca, and Daniele Pighin. 2016. Revisiting taxonomy induction over wikipedia. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2300–2309.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence 194:28–61.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. Artificial Intelligence 194:2–27.

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

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. 2014. Yago3: A knowledge base from multilingual wikipedias. In 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM 38(11):39–41.

Vivi Nastase, Michael Strube, Benjamin Börschinger, Cácilia Zirn, and Anas Elghafari. 2010. Wikinet: A very large scale multi-lingual concept network. In LREC. Citeseer.

Simone Paolo Ponzetto and Michael Strube. 2008. Wikitaxonomy: A large scale knowledge resource. In ECAI. Citeseer, volume 178, pages 751–752.

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

Paola Velardi, Stefano Faralli, and Roberto Navigli. 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&oldid=773693902. [Online; accessed 9-April-2017].

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Advances in neural information processing systems. pages 649–657.