Transforming Wikipedia into a Large-Scale Fine-Grained Entity Type Corpus

Abbas Ghaddar, Philippe Langlais
RALI-DIRO
Montreal, Canada
abbas.ghaddar@umontreal.ca, felipe@iro.umontreal.ca

Abstract

This paper presents WiFiNE, an English corpus annotated with fine-grained entity types. We propose simple but effective heuristics we applied to English Wikipedia to build a large, high quality, annotated corpus. We evaluate the impact of our corpus on the fine-grained entity typing system of Shimaoka et al. (2017), with 2 manually annotated benchmarks, FIGER (GOLD) and ONTONOTES. We report state-of-the-art performances, with a gain of 0.8 micro F1 score on the former dataset and a gain of 2.7 macro F1 score on the latter one, despite the fact that we employ the same quantity of training data used in previous works. We make our corpus available as a resource for future works.

Keywords: Annotated Corpus, Fine-Grained Entity Type, Wikipedia

1. Introduction

Entity typing is the task of classifying textual mentions into their respective types. While the standard Named-Entity Recognition (NER) task focuses on a small set of types (e.g. 4 classes defined by the CONLL shared task-2003 (Tjong Kim Sang and De Meulder, 2003)), fine-grained tagging deals with much larger type sets (e.g. 112 types used in (Ling and Weld, 2012)). Entity typing has received an increasing interest lately from the NLP community, due its role in Relation Extraction, Entity Linking, Question Answering, etc.

One issue in fine-grained typing is the absence of a well-established training corpus. The large number of types makes it difficult to manually annotate the amount of data needed for training. This bottleneck was addressed by using an automatic annotation procedure (Section 2), which follows two steps:

1. Identifying and linking entity mentions to a Knowledge Base (typically Freebase).
2. Assigning to each mention the set of types that apply in the context of the sentence.

Step 1 suffers a number of issues: lack of coverage when Wikipedia is used as a source (Ghaddar and Langlais, 2016b), and entity linking which is error prone (Ren et al., 2016). Step 2 also has limitations: the type of a mention is often resolved with strict pruning heuristics (regardless of the context) as in (Gillick et al., 2014); or the type of a mention is kept ambiguous following (Shimaoka et al., 2017). For instance, in the sentence: "*Gonzales* embarked on a pop career as the leader of the alternative rock band Son." The entity *Chilly Gonzales* has 3 labels in Freebase: *musician*, *writer* , *actor* but only *musician* applies here.

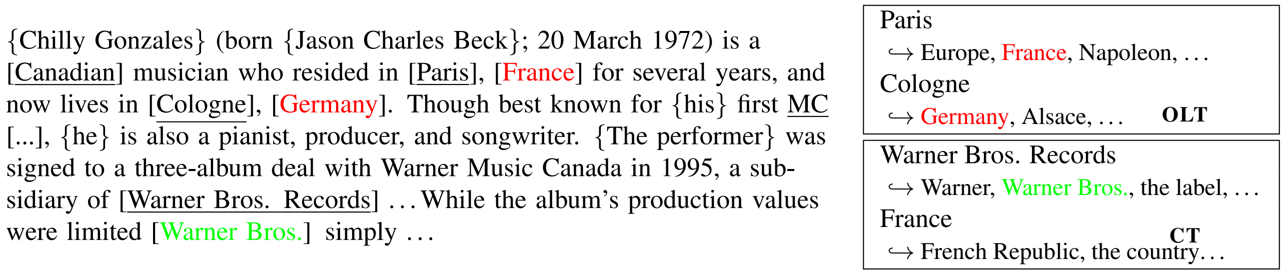
In this paper, we revisit the idea of automatically extracting fine-grained entity annotations from Wikipedia. Similarly to previous works, we gather annotations from anchored texts in an article, as well as their associated types in Freebase (Bollacker et al., 2008). In addition, we also generate annotations for texts not anchored in Wikipedia following (Ghaddar and Langlais, 2017). We do this by considering coreference mentions of anchored texts as candidate annotations, and by exploiting the out-link structure of Wikipedia. We propose an easy-first annotation pipeline described in Section 3 which happens to reduce noise. Second, we define simple yet efficient heuristics in order to prune the set of candidate types of each entity mention found in the article. These heuristics are based on: Freebase tuples, the high density of entity mentions, and the paragraph and section structure of the article.

We applied our methodology on a 2013 English Wikipedia dump, leading to a large annotated corpus called WiFiNE, which contains more annotations than similar corpora. We evaluate annotation quality intrinsically on a set of manually annotated mentions. We perform an extrinsic evaluation by training the entity typing model of (Shimaoka et al., 2017) on randomly generated subsets of WiFiNE. We compare the performances obtained by the resulting models on two well-established test sets: FIGER (GOLD) (Ling and Weld, 2012) and ONTONOTES (Gillick et al., 2014). The newly trained models clearly outperform previous ones on both benchmarks, demonstrating the superiority of our approach.

In summary, our contributions are the following:

- We provide over 110M proper name, nominal, and pronominal mentions annotated with fine-grained entity types in two taxonomies.
- We measure the efficiency of WiFiNE for training fine-grained entity typing. We outperform state-of the art results by 0.3 strict, and 0.8 macro F1 scores on the FIGER benchmark and by 0.9 strict, and 2.3 macro F1 scores on the OntoNotes dataset.

The remainder of paper is organized as follows. Section 2, discusses recent related works. We describe the annotation process along with the main statistics of our corpus in Section 3. An evaluation of WiFiNE on entity typing is described in Section 4, before concluding and discussing future works in Section 5.



{Chilly Gonzales} (born {Jason Charles Beck}; 20 March 1972) is a [Canadian] musician who resided in [Paris], [France] for several years, and now lives in [Cologne], [Germany]. Though best known for {his} first MC [...], {he} is also a pianist, producer, and songwriter. {The performer} was signed to a three-album deal with Warner Music Canada in 1995, a subsidiary of [Warner Bros. Records] ... While the album's production values were limited [Warner Bros.] simply ...

Paris
↪ Europe, France, Napoleon, ...
Cologne
↪ Germany, Alsace, ...
OLT
Warner Bros. Records
↪ Warner, Warner Bros., the label, ...
France
↪ French Republic, the country...
CT

Figure 1: Illustration of the process with which we gather annotations into WiFiNE for the target page https://en.wikipedia.org/wiki/Chilly_Gonzales. Square Bracketed segments are the annotations; curly brackets indicate mentions from the resource of (Ghaddar and Langlais, 2016a); while underlined text are anchored texts in the corresponding Wikipedia page. OLT represents the out-link table (which is compiled from the Wikipedia out-link graph structure), and CT represents the coreference table we gathered from the resource.

2. Related Works

In previous works, the entity mention detection process is performed using one of two methods. The first one consists in using the internal links in Wikipedia as training data, where anchored strings (that have an equivalent page in Freebase) are treated as entity mentions (Ling and Weld, 2012; Ren et al., 2016). Another method is to directly use a Freebase entity resolver such as DB-pedia Spotligh (Daiber et al., 2013) to link textual mentions to their Freebase page (Gillick et al., 2014; Yogatama et al., 2015; Ren et al., 2016).

In both cases, the Freebase object_type attributes of the entity are mapped to a predefined set of types. In the last few years, two popular mapping schemes emerged: FIGER (Ling and Weld, 2012) (112 label) and GILLICK (Gillick et al., 2014) (89 label). They are both organized in a hierarchical structure, where children labels also inherit the parent label. FIGER defines a 2-level hierarchy (e.g. /person and /person/musician); while GILLICK uses 3 levels of types (e.g. /person and /person/artist, /person/artist/musician).

Most resolved entities have multiple type labels, but not all of them typically apply in a given context. One solution consists in ignoring the issue, and instead relying on the robustness of the model to deal with heterogeneous labels; this approach is adopted by (Yogatama et al., 2015; Shimaoka et al., 2017). Another solution involves filtering. In (Ling and Weld, 2012; Gillick et al., 2014), the authors apply hard pruning heuristics:

- **Sibling pruning** Removes sibling types if they came from a single parent type. For instance, a mention labelled as /person/artist/musician and /person/artist/actor would be tagged by /person/artist and /person.
- **Minimum count pruning** All labels that appear once in the document are removed. For example, if multiple entities in a document are labelled as /person/artist/musician and only one of them have /person/artist/actor as an extra label, the latter is considered noisy.

Such heuristics decrease the number of training data by 40-45% according to (Gillick et al., 2014; Ren et al., 2016). Ren et al. (2016) propose a distant supervision approach to deal with noisy labelled data. Their method consists in using unambiguous mentions to de-noise mentions with heterogeneous labels that appear in a similar context.

Because only a tiny portion of texts in Wikipedia are anchored, some strategies are needed to infer more annotations. In this study, we revisited the approach of (Ghaddar and Langlais, 2017) which consist in annotating Wikipedia with coarse-grained entity type (PER, LOC, ORG and MISC), resulting in a corpus called WiNER. In this paper, we propose to extend this approach with more types and mentions, leading to WiFiNE. First, we enrich the corpus with nominal and pronominal coreference mentions, then we extend the set of types (4 previously) to either 112 (FIGER) or 89 (GILLICK). In the next Section, we summarize the original process proposed by (Ghaddar and Langlais, 2017) and then we describe our extensions.

3. WiFiNE

3.1 Mention Recognition

The pipeline used to extract annotations from Wikipedia is illustrated in Figure 1, for an excerpt of the Wikipedia article Chilly_Gonzales, hereafter named the target article. The anchored texts of out-links in the target article are elected entity mentions. For instance, we identify *Warner Bros. Records* and *Paris* as mentions. In general, a Wikipedia article has an equivalent page in Freebase. We remove mentions that do not have such a page. This way, we filter out anchored texts that are not named-entities (such as *List of Presidents of the United States*).

Because the number of anchored strings in Wikipedia is rather small — less than 3% of the text tokens — we propose to leverage: (1) the out-link structure of Wikipedia, (2) the information of all the surface strings used to describe the main concept of a Wikipedia article. For the latter, we rely on the resource[1] described in (Ghaddar and Langlais, 2016a) that lists, for all the articles in Wikipedia (those that have a Freebase counterpart), all the text mentions that are coreferring to the main concept of an article (CT of Figure 1). For instance, for the article *Chilly_Gonzales*, the

[1] http://rali.iro.umontreal.ca/rali/en/wikipedia-main-concept

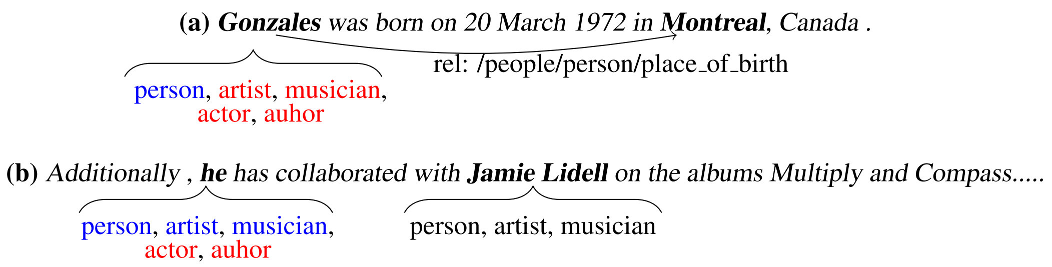

(a) Gonzales was born on 20 March 1972 in Montreal, Canada .
rel: /people/person/place_of_birth
person, artist, musician,
actor, auhor
(b) Additionally , he has collaborated with Jamie Lidell on the albums Multiply and Compass.....
person, artist, musician,
actor, auhor
person, artist, musician

Figure 2: Illustration of our de-noise heuristic rules. Spans in bold are entity mentions. Blue labels are relevant ones, while red ones are irrelevant.

resource lists proper names (e.g. Gonzales, Beck), nominal (e.g. the performer) and pronominal (e.g. he) mentions that refer to Chilly Gonzales. Our strategy for collecting extra annotations is a 4-step process, where:

1. We consider direct out-links of the target article. We search the titles of the articles we reach that way. We also search for their coreferences as listed in the resource of (Ghaddar and Langlais, 2016a). For instance, we search (exact match) *Warner Bros. Records* and its coreferences (e.g. *Warner*, *Warner Bros.*) in the target article.
2. We follow out-links of out-links, and search in the target article (by an exact string match) the titles of the articles reached. For instance, we search for the strings *Europe*, *France*, *Napoleon*, as well as other article titles from the out-link list of the article *Paris*.
3. For the titles matched during step 2, we also match their coreferent mentions. For instance, because *France* was matched in the previous step, we also search its coreferences as listed in the coreference table (CT of Figure 1).
4. Last, we adapt the multi-sieve rule-based coreference resolver of (Raghunathan et al., 2010) to the specificity of Wikipedia in order to find the antecedent referents of a pronominal mention. The rules link a pronoun to its best antecedent mention based on attributes agreement (gender, number, entity type,...). We apply the pronoun coreference rules on each article, then discard all pronouns that do not refer to a Wikipedia entity mention.

During this process, some collisions may occur. We solve the issue of overlapping annotations by applying the steps exactly in the order presented above. Our steps have been ordered in such a way that the earlier the step, the more confidence we have in the strings matched at that step. It may also happen that two out-link articles contain the same mention (for instance *Washington_State* and *George_Washington* both contain the mention *Washington*), in which case we annotate this ambiguous mention with the type of the closest[2] unambiguous mention.

3.2 Manual Evaluation

Step 1 raises the coverage from less than 3% to 9.5%, step 2 further raises it to 11.5%, while step 3 and 4 increase it to 23% and 30% respectively. We assessed the annotation quality of a random subset of 1000 mentions. While we measure an accuracy of 92% and 88% for mentions detected during step 1 and 2 respectively, the accuracy decreases to 81% and 77% during step 3 and 4 respectively. We identified two main sources for errors in the coreferent mentions detection procedure.

a) *[Eldridge Pope]* was a traditional brewery.....Sixteen years later the *[Pope]*$_\star$ brothers floated the business...

b) Montreal Impact's biggest rival is *[Toronto FC]* because Canada's two largest cities have rivalries in and out of sport. Montreal and *[Toronto]*$_\star$ professional soccer teams have competed against each other for over 40 years.

Figure 3: Examples of errors in our annotation pipeline. Faulty annotations are marked with a star.

One source of error comes from the resource used to identify the mentions of the main concept. We measured in a previous work (Ghaddar and Langlais, 2016a), that the process we rely on for this (a binary classifier) has an accuracy of 89%. Example (a) of Figure 3 illustrates such a mistake where the family name *Pope* is wrongly assumed coreferent to the brewery *Eldridge Pope*. We also found that our 4-step process and the disambiguation rules fail in 15% of the cases. Figure 3 b) illustrates an example where we erroneously recognize the mention *Toronto* (referring to the town) as a coreferent of the (non ambiguous mention) *Toronto FC*, simply because the latter is close to the former.

3.3 Type Mapping

Following previous works, we map Freebase `object_type` attributes of each entity mention detected to a set of fine-grained types. An entity mention is said to be clean if its labels belong to only a single path (not necessarily a leaf); otherwise, it is noisy. For example, the mentions *France* or *Germany* with labels `/location` and `/location/country` are considered clean. On the other hand, the entity mention *Chilly Gonzales* annotated with 5 labels (`/person`, `/person/artist`, `/person/artist/musician`, `/person/artist/actor`, and `/person/artist /author`) is considered noisy because only one of the

[2] Before or after the named-entity.

last three types is qualified in a given context (see Fig. 2). We measured that 23% of mentions in WiFiNE that have two labels or more don't belong to a single path (noisy), and 47% of those have more than 2 noisy labels (e.g. *Gonzales* in Fig. 2). We propose to eliminate noisy labels in WiFiNE using rules based on the high coverage of entity mentions, coupled with Freebase triples and the paragraph and section structure of Wikipedia:

1. **Freebase Relation Type:** We label the mention by the type indicated by the relation. A Freebase relation is a concatenation of a series of fragments. The first two fragments of the relation indicate the Freebase type of the subject, and the third fragment indicates the relation type. In example (a) of Fig. 2, the triple (arg1: *Chilly Gonzales*; rel: `/people/person/place_of_birth`; arg2: *Montreal*) found in Freebase indicates that only `/person` should apply to the *Gonzales* mention in this context.
2. **Common Attribute Sharing:** If a non-ambiguous mention (*Jamie Lidell* in example (b)) has a type set which is a subset of another mention with noisy labels (*he*, referent of *Chilly Gonzales*) occurs in the same sentence, we assign to the noisy mention the common labels between both mentions.

We first apply our rules at the sentence level, then at the paragraph and section level. Whenever we de-noise an entity mention in such a way, all its coreferent mentions (in the scope) receive the same type.

Heuristic	**Pre**	**Rec**	**F1**
w/o Rules	31.8	**100.0**	48.3
Rule-1 only	48.8	87.2	62.3
Rule-2 only	56.4	85.6	68.0
Both Rules	**79.2**	81.8	**80.5**
Level of Application			
Sentence	66.5	85.5	73.7
+ Paragraph	72.7	82.6	78.6
+ Section	**79.2**	81.8	**80.5**

Table 1: De-noising rules evaluation on 1000 hand-labelled mentions following GILLICK type hierarchy.

We assessed the quality of our de-noising rules on 1000 randomly selected noisy mentions. Table 1 reports precision, recall and F1 scores on the ablation study of the proposed heuristics. We start with an accuracy of 48% when either rule is applied. We measure performance after removing labels identified as noisy by rule one, two and both. Also, we measure the accuracy when the rules are applied at sentence, paragraph and section levels. Results show that our rules greatly improve the annotation quality by roughly 32%. Also, we observe that the first rule is more important than the second, but both rules complement each other. As expected, applying the rules at paragraph and section levels further improve the performance. We identify two sources of errors: (1) pruning heuristics don't apply to 11% of mentions; (2) our rules failed to pick up the correct label in 9% of the cases. Example (a) of Figure 4 illustrates such a mistake where *Gonzales* is labelled as *musician* rather than *author* because *Feist* is considered as *musician* in this context. In example (b), *Gonzales* is wrongly labelled as *person* thought that the relation `/people/person/nationality` exist between both entity but the sentence don't state it.

a) *[Gonzales]*$_{musician\star}$ returned as a contributor on *[Feist]*'s 2007 album...

b) *[Gonzales]*$_{person\star}$ said in an interview: My experiences in *[Canada]* had been disappointing

Figure 4: Examples of errors in our de-noising rules. Faulty annotations are marked with a star.

Table 2 illustrates a randomly-picked selection of mentions annotated in WiFiNE, along with their type according to the GILLICK scheme. The last two examples illustrate noisy annotations. In the first one our process failed to distinguish between the company and its product. The second example is a mention detection error, we couldn't recognize *Viitorul Homocea* as an entity, because this soccer team does not have a page in Wikipedia or Freebase.

3.4 Corpus Statistics

WiFiNE is built from 3.2M Wikipedia articles, comprising more than 1.3G tokens accounting for 54M sentences, 41M of which contain at least one entity mention. Overall, it gathers 182.7M mentions: 95.1M proper, 62.4M nominal and 24.2M pronominal ones. Table 3 summarizes the mention statistics and label distribution over the number of levels of FIGER and GILLICK type hierarchies.

	FIGER	**GILLICK**
Total mentions	159.4	111.1
Proper mentions	82.5 (52%)	64.8 (58%)
Nominal mentions	55.9 (35%)	29.8 (27%)
Pronominal mentions	21.0 (13%)	16.5 (15%)
Total Labels	243.2	230.9
Level 1	153.8 (63%)	111.1 (48%)
Level 2	89.5 (37%)	90.0 (39%)
Level 3	-	29.8 (13%)

Table 3: Mention statistics and label distribution (in millions and percentages) over the number of levels of FIGER and GILLICK type hierarchy.

First, we note that the total number of mentions in FIGER and GILLICK is less than the total number of entity mentions. This is because: (a) we remove noisy mentions that our rules failed to disambiguate (11%), (b) some mentions cannot be mapped to either schemes (e.g. fictional characters). Second, we note that FIGER mentions out number those of GILLICK, simply because their scheme covers more types (112 vs 89).

Sentence	Labels
In **Kent v. Dulles** , 357 U.S. 116 (1958) , the Court held that the federal government . . .	/other /other/event
The **Cangrejal River** or **Río Cangrejal** is a river that drains several mountain tributaries . . .	/location /location/geography /location/geography/body_of_water
. . . editions of **Millionaire** to be aired between 7:00 and 7:30 pm	/other /other/art /other/art/broadcast
Mies Bouwman stopped **her** regular work after falling sick but has occasionally	/person /person/artist
. . . to imprisoned Christians and niece of the **Emperor Gallienus** , found Anthimus in prison .	/person /person/political_figure
. . . of **vinyl siding** which does not weather as wood does .	/other /other/product
The firm was the first state-owned rail vehicle in Argentina. . .	/organization /organization/company
The **1 – 2 ton** was a sailing event on the **Sailing at the 1900 Summer Olympics** program in Meulan	/other /other/event /other/event/sports_event
He took part in the White Council after **Sauron** 's return.. . .	/person /person/artist /person/artist/actor
Clove is **Syzygium aromaticum** and belongs to division of **Magnoliophyta** in the kingdom Plantae .	/other /other/living_thing
Pepsi⋆ also created a fellowship at Harvard University which enable students from. . .	/other /other/food
. . . Viitorul **Homocea**⋆ , Siretul Suraia and Trotusul Ruginesti deducted 3 points .	/location

Table 2: Random selection of annotations from WiFiNE following GILLICK type hierarchy. Faulty annotations are marked with a star.

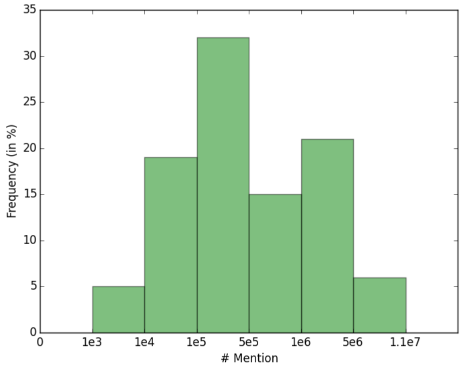


Figure 5: Distribution of entity type labels according to the FIGER type hierarchy.

Following the GILLICK scheme, each mention has 2 types on average, where 39% of them are of level 2, and 13% are of level 3. The distribution of level 2 and 3 labels in WiFiNE exceed its equivalent in the ONTONOTES (Gillick et al., 2014) dataset (29% and 3% respectively). Figure 5 illustrates the percentage of types that recieve a given number of mentions in WiFiNE. It shows that the majority of types have more than 100k mentions and roughly 25% (like `city`, `company`, `date`) exceeds 1M mentions. Also, we observe that 5% of the types have less than 10k mentions (e.g. `/event/terrorist_attack`), and none of them has less than 1k mentions[3].

4. Evaluation on Entity Typing

4.1 Reference System

In all experiments, we deploy the off the shelf neural network model of (Shimaoka et al., 2017). Given a mention in its context, the model uses three representations in order to associate the mention with the correct types.

- **Mention representation:** the average of the mention words embedding.
- **Context representation:** First, a Bi-LSTM model is applied on the left and right context of the mention, then an attention layer is placed on top of the model.
- **Feature Representation:** They learn the representations of hand-crafted features.

We trained the tagger on various subsets of WiFiNE as described in the next section. We use the default configuration of the tagger, except the batch size which we set to 100 rather than 1000 and the learning rate that we changed from 0.001 to 0.0005[4]

4.2 Datasets and Evaluation Metrics

We evaluate the model on two manually annotated benchmark: FIGER (GOLD) (Ling and Weld, 2012) and ONTONOTES (Gillick et al., 2014). The first consist of 18 news reports annotated following FIGER scheme, while the second are 77 documents from the OntoNotes 5.0 (Weischedel et al., 2013) test set annotated according to the GILLICK scheme. Following previous works, we used Strict, loose Macro-averaged, and loose Micro-averaged

[3] A similar distribution is obtained with GILLICK type hierarchy.

[4] We observed better results on the held-out development set.

F1 scores as metrics for evaluation. Strict measures exact match, while losses metrics measure macro/micro partial matches between gold and system labels. Macro is the average of F1 scores on all types, while Micro is the harmonic mean. Table 4 and 6 compared the performance obtained by the resulting models with those of previous works on FIGER (GOLD) and ONTONOTES test set respectively. We perform an ablation test on our 4-step process of Section 3.1 by training the model on 7 variants of WiFiNE:

- **Line 1-3:** hyperlinks + proper name coreference mentions (step 1 and 2 of Section 3.1)
- **Line 4:** hyperlinks + proper name + nominal coreference mentions (step 1-3 of Section 3.1).
- **Line 5:** hyperlinks + proper name + pronominal coreference mentions (step 1, 2 and 4 of Section 3.1).
- **Line 6-7:** hyperlinks + proper name + nominal + pronominal coreference mentions (all steps).

The goal is to validate if proper name, nominal and pronominal coreference mentions are necessary to fine-grained entity tying performance. For each variant, we report the average score on 5 randomly generated subsets. To be comparable with previous works, we used training materiel up to 4 million mentions, and leave experiments on the usefulness of the full WiFiNE for future work.

4.3 Results on FIGER (GOLD)

Previous works trained their models on 2.6 million mentions obtained by mapping hyperlinks in Wikipedia articles to Freebase[5].

Models				Strict	Macro	Micro
FIGER (Ling and Weld, 2012)				52.30	69.90	69.30
FIGER+PLE (Ren et al., 2016)				59.90	76.30	74.90
Attentive (Shimaoka et al., 2017)				59.68	78.97	75.36
(Abhishek et al., 2017)				65.80	**81.20**	77.40
	Proper	**Nominal**	**Pronominal**	**This work**		
(1)	1	0	0	61.99	76.20	75.12
(2)	2	0	0	63.77	77.56	76.25
(3)	3	0	0	63.41	78.03	76.32
(4)	1	1	0	64.83	79.26	77.36
(5)	1	0	1	63.06	79.00	76.77
(6)	1	1	1	65.19	79.59	77.55
(7)	2	1	1	**66.07**	79.94	**78.21**

Table 4: Results of the reference system trained on various subsets of WiFiNE, compared to other published results on the FIGER (GOLD) test set. Training data (in millions) include: proper name; nominal and pronominal mentions.

Our model trained on 4M mention (line 7) outperforms the initial model of (Shimaoka et al., 2017) by 6.2, 1.0 and 2.9 on strict, micro, macro F1 scores, and the state-of-the-art of (Abhishek et al., 2017) by 0.3 and 0.9 strict and macro F1 scores. First, we observe that using hyperlinks and proper name mentions (line 3) for training improves the performance of the original model of (Shimaoka et al., 2017) that uses data driven from hyperlinks only. Second, we notice that models trained on a mix of proper name and nominal (line 4) or pronominal (line 5) coreference mentions outperform the model trained on proper name mentions (line 2) solely. Third, we observe that the combination of 3 mention types (line 6-7) is required in order to outperform the state-of-the-art, which validate our 4-step method of Section 3.1.

Label type	FIGER (GOLD)	WiFiNE
/person	31.5%	16.6%
/organization	16.9%	7.7%
/location	13.2%	13.6%
/location/city	5.0%	4.3%
/organization/sports_team	4%	1.0%

Table 5: Comparison of the distribution of the top 5 types present in FIGER (GOLD) test set to that of WiFiNE.

Table 5 shows the 5 most frequent types the FIGER (GOLD) test set compared to those in WiFiNE. FIGER (GOLD) is a small dataset, it contains only 523 mentions annotated with 41 different labels. We observe that the type distribution in this dataset follows a zipfian curve, while the distribution of types in WiFiNE is similar to a normal distribution (Figure 5). Figure 6 illustrates some errors committed on FIGER (GOLD) dataset. Error mostly occur on mentions with labels that don't belong to a single path (example a), and on ambiguous mentions (example b).

(a) . . . bring food for the employees at [Safeway] . . .
Gold: /location /location/city
/organization /organization/company
Pred: /organization /organization/company

(b) *With the huge popularity of [EyeFi] cards . . .*
Gold: /product
Pred: /organization

Figure 6: Examples of mentions erroneously classified in FIGER (GOLD) dataset.

4.4 Results on OntoNotes

Ren et al. (2016), Shimaoka et al. (2017) and Abhishek et al. (2017) trained their models on newswire documents present in OntoNotes (Weischedel et al., 2013), where entity mentions were automatically identified and linked to Freebase using `DB-pedia Spotligh` (Daiber et al., 2013). On the other hand, Gillick et al. (2014) and Yogatama et al. (2015) used an entity linker to automatically annotated 113k news documents. Results on the ONTONOTES dataset validate the observation we obtained on FIGER (GOLD). Models trained on proper names in addition to nominal (line 4 in Table 6) or pronominal (line 5) coreference mentions is better than only training on proper names (line 2). In addition, training on the combination of all coreference mentions (line 6-7) systematically improves performances.

[5] The dataset is distributed by (Ren et al., 2016)

Models				Strict	Macro	Micro
(Gillick et al., 2014)				N/A	N/A	70.0
K-WASABIE (Yogatama et al., 2015)				N/A	N/A	**72.98**
FIGER+PLE (Ren et al., 2016)				57.20	71.50	66.10
Attentive (Shimaoka et al., 2017)				51.74	70.98	64.91
(Abhishek et al., 2017)				52.20	68.50	63.30
	Proper	**Nominal**	**Pronominal**	**This work**		
(1)	1	0	0	55.25	68.21	61.49
(2)	2	0	0	57.05	71.96	66.03
(3)	3	0	0	57.47	72.87	66.97
(4)	1	1	0	57.17	73.07	67.30
(5)	1	0	1	57.50	73.08	67.35
(6)	1	1	1	57.80	73.60	67.82
(7)	2	1	1	**58.05**	**73.72**	67.97

Table 6: Results of the reference system trained on various subsets of WiFiNE, compared to other published results on the ONTONOTES test set. Training data (in millions) includes: proper name; nominal and pronominal mentions.

We outperform best results reported by previous works on strict, macro F1 scores by 0.9 and 2.3 receptively. On the other hand, we underperform (Gillick et al., 2014) and (Yogatama et al., 2015) and by 3 and 5 point on the micro metric respectively. In (Gillick et al., 2014; Yogatama et al., 2015), the authors do not report results on strict and macro metrics and neither their models nor their training data are available. Consequently, we couldn't specify the cause of the gap on the micro metric, but we report some improvement over (Shimaoka et al., 2017) model on the loose metrics. A potential reason behind this gap is that the text genre of their training data and that of ONTONOTES is the same (newswire). Our models were trained on randomly picked Wikipedia sentences (out of domain). Also, we note that in order to generate their corpus, (Gillick et al., 2014; Yogatama et al., 2015) applied filtering rules that are responsible for the loss of 45% of the mentions. We have no such heuristic here, but we still observe competitive performances.

Label type	Onto Test	WiFiNE
`/other`	44.0%	20.0 %
`/organization`	10.5%	6.3 %
`/person`	8.4%	17.6%
`/organization/company`	7.7%	2.3%
`/location`	7.6%	11.8%

Table 7: Comparison of the distribution of the top 5 types present in ONTONOTES test set to that of WiFiNE.

Table 7 shows the 5 most frequent types in the ONTONOTES dataset and in WiFiNE. Although ONTONOTES is much larger the FIGER (GOLD)[6], we still observe that the distribution of types in this dataset is zipfian. We also note that the type `/other` is over-represented (44%) in this dataset, because Gillick et al. (2014) annotated all non-entity mentions (examples in table 8) as `/other`. We observe that 73% of the wrong decisions that our model made on ONTONOTES are committed on this type. In WiFiNE, `/other` always refers to an entity mention, and in most cases the mention has an additional level two and three labels.

trouble
addition
personal reasons
some complications
additional evidence
diplomatic relations
a modest pretax gain
the active role taken
in the affairs of United
quotas on various economic indicators
the invitation of the Foreign Affairs Institute
amounts related to areas where deposits are received

Table 8: Examples of non-entity mentions annotated as `/other` in the of OntoNotes test set.

5. Conclusion

We built on the work of (Ghaddar and Langlais, 2017) which developed WiNER, a coarse-grained entity type corpus made merly from English Wikipedia articles, and propose WiFiNE, a fine-grained entity type corpus annotated with nominal and pronominal coreference mentions. We evaluated the impact of our corpus on a neural network tagging system with 2 human made benchmarks. Experiments shows state-of-the-art performances on both benchmarks, when WiFiNE is used as training materiel. Our analysis on both datasets indicates the following observations. First, enriching Wikipedia articles with proper names, nominal and pronominal mentions systematically leads to better performances, which validate our 4-step approach. Second, the correlation between the train and test type distribution is an important factor to entity typing performance. Third, models could benefit from an example selection strategy based on the genre of the test set. As future work, we want to study the usefulness of WiFiNE on a NER in Tweets, and if models can benefits from the full corpus. WiFiNE is publicly available at `http://rali.iro.umontreal.ca/rali/en/wifiner-wikipedia-for-et`. We hope this resource will foster further research on fine-grained entity type tagging.

6. Acknowledgements

This work has been founded by TRIBE, a Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Training Experience (CREATE) Program. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. We thank the anonymous reviewers for their insightful comments.

[6] It contains roughly 9000 mentions annotated with 88 different types

7. Bibliographical References

Abhishek, A., Anand, A., and Awekar, A. (2017). Fine-Grained Entity Type Classification by Jointly Learning Representations and Label Embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 797–807, Valencia, Spain, April. Association for Computational Linguistics.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Ghaddar, A. and Langlais, P. (2016a). Coreference in Wikipedia: Main Concept Resolution. In *CoNLL*, pages 229–238.

Ghaddar, A. and Langlais, P. (2016b). WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 05/2016.

Ghaddar, A. and Langlais, P. (2017). WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 413–422.

Gillick, D., Lazic, N., Ganchev, K., Kirchner, J., and Huynh, D. (2014). Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*.

Ling, X. and Weld, D. S. (2012). Fine-Grained Entity Recognition. In *AAAI*.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Ren, X., He, W., Qu, M., Voss, C. R., Ji, H., and Han, J. (2016). Label noise reduction in entity typing by heterogeneous partial-label embedding. *arXiv preprint arXiv:1602.05307*.

Shimaoka, S., Stenetorp, P., Inui, K., and Riedel, S. (2017). Neural Architectures for Fine-grained Entity Type Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1271–1280, Valencia, Spain, April. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

Yogatama, D., Gillick, D., and Lazic, N. (2015). Embedding Methods for Fine Grained Entity Type Classification. In *ACL (2)*, pages 291–296.