How Bad are PoS Taggers in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project

Guillaume Wisniewski and **François Yvon**
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France
{guillaume.wisniewski,francois.yvon}@limsi.fr

Abstract

The performance of Part-of-Speech tagging varies significantly across the treebanks of the Universal Dependencies project. This work points out that these variations may result from divergences between the annotation of train and test sets. We show how the annotation variation principle, introduced by Dickinson and Meurers (2003) to automatically detect errors in gold standard, can be used to identify inconsistencies between annotations ; we also evaluate their impact on prediction performance.

1 Introduction

The performance of Part-of-Speech (PoS) taggers significantly degrades when they are applied to test sentences that depart from training data. To illustrate this claim, Table 1 reports the error rate achieved by our in-house PoS tagger on the different combinations of train and test sets of the French treebanks of the Universal Dependencies (UD) project (Nivre et al., 2018). [1] It shows that depending on the train and test sets considered, the performance can vary by a factor of more than 25.

Many studies (Foster, 2010; Plank et al., 2014) attribute this drop in accuracy to *covariate shift* (Shimodaira, 2000), characterizing the differences between domains by a change in the marginal distribution $p(\mathbf{x})$ of the input (e.g. increase of out-of-vocabulary words, missing capitalization, different usage of punctuation, etc), while assuming that the conditional label distribution remains unaffected.

This work adopts a different point of view : we believe that the variation in tagging performance is due to a *dataset shift* (Candela et al., 2009), i.e. a change in the joint distribution of the features and labels. We assume that this change mainly results from incoherences in the annotations between corpora or even within the same corpus. Indeed, ensuring inter-annotator agreement in PoS tagging is known to be a difficult task as annotation guidelines are not always interpreted in a consistent manner (Marcus et al., 1993). For instance, Manning (2011) shows that many errors in the WSJ corpus are just mistakes rather than uncertainties or difficulties in the task ; Table 2 reports some of these *annotation divergences* that can be found in UD project. The situation is naturally worse in cross-corpora settings, in which treebanks are annotated by different laboratories or groups.

The contribution of this paper is threefold :

— we show that, as already pointed out by de Marneffe et al. (2017), the variation principle of Boyd et al. (2008) can be used to flag potential annotation discrepancies in the UD project. Building on this principle, we introduce, to evaluate the annotation consistency of a corpus, several methods and metrics that can be used, during the annotation to improve the quality of the corpus.

— we generalize the conclusions of Manning (2011), highlighting how error rates in PoS tagging are stemming from the poor quality of annotations and inconsistencies in the resources ; we also systematically quantify the impact of annotation variation on PoS tagging performance for a large number of languages and corpora.

— we show that the evaluation of PoS taggers in cross-corpora settings (typically in domain adaptation experiments) is hindered by systematic annotation discrepancies between the corpora and quantify the impact of this divergence on PoS tagger evaluation. Our observations stress the fact that comparing in- and out-domain scores as many

1. See Section 2 for details regarding our experimental setting

Proceedings of NAACL-HLT 2019, pages 218–227
Minneapolis, Minnesota, June 2 - June 7, 2019.

test → ↓ train	FTB	GSD	ParTUT	SRCMF	Sequoia	Spoken	PUD
FTB	**2.8%**	7.0%	6.5%	45.4%	5.4%	18.7%	12.9%
GSD	6.7%	**3.7%**	7.2%	45.5%	5.4%	16.3%	10.2%
ParTUT	11.2%	10.9%	**5.9%**	55.7%	11.3%	22.9%	15.8%
SRCMF	38.8%	37.8%	36.2%	**7.5%**	37.4%	34.7%	36.1%
Sequoia	7.5%	7.5%	8.4%	48.0%	**4.0%**	19.3%	13.6%
Spoken	32.1%	30.3%	25.7%	51.8%	29.5%	**7.9%**	30.1%

Table 1: Error rate (%) achieved by a PoS tagger trained and tested on all possible combinations of the French train and test sets of the UD project. To mitigate the variability of our learning algorithm, all scores are averaged over 10 training sessions.

works do (e.g. to evaluate the quality of a domain adaptation method or the measure the difficulty of the domain adaptation task) can be flawed and that this metrics has to be corrected to take into account the annotation divergences that exists between corpora.

The rest of this paper is organized as follows. We first present the corpora and the tools used in our experiments (§ 2). We then describe the annotation variation principle of Dickinson and Meurers (2003) (§ 3) and its application to the treebanks of the Universal Dependencies project (§ 4). We eventually assess the impact of annotation variations on prediction performance (§ 5 and § 6).

The code and annotations of all experiments are available on the first author website. [2] For the sake of clarity, we have only reported our observations for the English treebanks of the UD project and, sometimes, for the French treebanks (because it has seven treebanks). Similar results have however been observed for other languages and corpora.

2 Experimental Setting

Data All experiments presented in this work use the Universal Dependencies (UD) 2.3 dataset (Nivre et al., 2018) that aims at developing cross-linguistically consistent treebank annotations for a wide array of languages.

This version of the UD project contains 129 treebanks covering 76 languages. Among those, 97 treebanks define a train set that contains between 19 sentences and 68,495 sentences and a test set that contains between 34 and 10,148 sentences. For 21 languages, several test sets are available : there are, for instance, 7 test sets for French, 6 for English, 5 for Czech and 4 for Swedish, Chinese, Japanese, Russian and Italian. Overall, it is possible to train and test 290 taggers (i.e. there are 290 possible combinations of a train and a test set of the same language), 191 of these *conditions* (i.e. pairs of a train set and a test set) correspond to a cross-corpus setting and can be considered for domain adaptation experiments.

Many of these corpora [3] result from an automatic transformation (with, for some of them, manual corrections) from existing dependency or constituent treebanks (Bosco et al., 2013; Lipenkova and Souček, 2014). Because most treebanks have been annotated and/or converted independently by different groups, [4] the risk of inconsistencies and errors in the application of annotation guidelines is increased. There may indeed be several sources of inconsistencies in the gold annotations : in addition to the divergences in the theoretical linguistic principles that governed the design of the original annotation guidelines, inconsistencies may also result from automatic (pre-)processing, human post-editing, or human annotation. Actually, several studies have recently pointed out that treebanks for the same language are not consistently annotated (Vilares and Gómez-Rodríguez, 2017; Aufrant et al., 2017). In a closely related context, Wisniewski et al. (2014) have also shown that, in spite of common annotation guidelines, one of the main bottleneck in cross-lingual transfer between UD corpora is the difference in the annotation conventions across treebanks and languages.

2. https://perso.limsi.fr/wisniews/recherche/#coherence

3. For PoS, only 23 treebanks have been manually annotated natively with the Universal PoS tagset.

4. almost 65% of the UD contributors have participated in the annotation of only one corpus; for more than 15% of the treebanks all contributors have annotated a single corpus.

①	◇ With regard to the effect of the programme **on the convergence of high level$_{\text{ADJ}}$ training for trainers ,** it was not possible to make an assessment as there was not sufficient information on the link between national strategies and the activities under Pericles . ◇ With a view to enabling the assessment of the effect of the programme , among others **on the convergence of high level$_{\text{NOUN}}$ training for trainers ,** the evaluator recommends the preparation of a strategy document , to be finalised before the new Pericles enters into effect .
②	◇ **Notice$_{\text{NOUN}}$ Regarding Privacy and Confidentiality : PaineWebber reserves the right to monitor and review the content of all e-mail communications sent and or received by its employees .** ◇ **Notice$_{\text{PROPN}}$ Regarding Privacy and Confidentiality : PaineWebber reserves the right to monitor and review the content of all e-mail communications sent and or received by its employees .**
③	◇ The above applies to the Work as incorporated in a Collective Work , but this does not require the Collective Work apart from the **Work itself to be made subject$_{\text{ADJ}}$ to the terms of this License.** ◇ The above applies to the Derivative Work as incorporated in a Collective Work , but this does not require the Collective Work apart from the Derivative **Work itself to be made subject$_{\text{NOUN}}$ to the terms of this License .**

Table 2: Examples of annotation divergences in the English Web Treebank (EWT) corpus : these sentences share some common words (in bold) that do not have the same annotation. Only the labels that differ are represented.

PoS tagger In all our experiments, we use a history-based model (Black et al., 1992) with a LaSO-like training method (Daumé III and Marcu, 2005). This model reduces PoS tagging to a sequence of multi-class classification problems : the PoS of the words in the sentence are predicted one after the other using an averaged perceptron. We consider the standard feature set for PoS tagging (Zhang and Nivre, 2011) : current word, two previous and following words, the previous two predicted labels, etc. This 'standard' feature set has been designed for English and has not been adapted to the other languages considered in our experiments.

Our PoS tagger achieves an average precision of 91.10% over all UD treebanks, a result comparable to the performance of `UDPipe` 1.2 (Straka and Straková, 2017), the baseline of CoNLL'17 Shared Task '*Multilingual Parsing from Raw Text to Universal Dependencies*' that achieves an average precision of 91.22%. When not otherwise specified, all PoS tagging scores reported below are averaged over 10 runs (i.e. independent training of a model and evaluation of the test performance).

3 Annotation variation principle

The *annotation variation principle* (Boyd et al., 2008) states that if two identical sequences appear with different annotations, one of these two label sequences may be inconsistently annotated. Our work relies on this principle to identify discrepancies in the PoS annotation of treebanks.

We call *repeat* a sequence of words that appears in, at least, two sentences and *suspicious repeat* a repeat that is annotated in at least two different ways. Identifying suspicious repeats requires, first, to find all sequences of words that appear in two different sentences ; this is an instance of the *maximal repeat problem* : a *maximal repeat*, is a substring that occurs at least in two different sentences and cannot be extended to the left or to right to a longer common substring. Extracting maximal repeats allows us to find all sequence of words common to at least two sentences without extracting all their substrings. This problem can be solved efficiently using Generalized Suffix Tree (GST) (Gusfield, 1997) : if the corpus contains n words, extracting all the maximal repeats takes $\mathcal{O}(n)$ to build the GST and $\mathcal{O}(n)$ to list all the repeats. PoS annotations for these repeats can then be easily extracted and the ones that are identical can be filtered out to gather all suspicious repeats in a set of corpora. A detailed description of our implementation can be found in (Wisniewski, 2018).

Filtering heuristics Suspicious repeats can of course correspond to words or structures that are

ambiguity	◇ The early voting suggests that this time the Latin Americans **will come out to**$_{\textbf{PART}}$ vote in greater numbers , but it is unclear whether the increase will have an impact . ◇ Keep his cage open and go on your computer , or read a book , etc and maybe he **will come out to**$_{\textbf{ADP}}$ you .
inconsistency	◇ Trudeau will extend that invitation to the 45th **president**$_{\textbf{NOUN}}$ **of the United**$_{\textbf{ADJ}}$ **States**$_{\textbf{NOUN}}$, whoever he or she may be . ◇ I am GEORGE WALKER BUSH , son of the former **president**$_{\textbf{PROPN}}$ **of the United**$_{\textbf{PROPN}}$ **States**$_{\textbf{PROPN}}$ of America George Herbert Walker Bush , and currently serving as President of the United States of America .

Table 3: Example of an actual ambiguity and of an annotation inconsistency between the English EWT and PUD corpora. Repeated words are in bold and words with different PoS in red.

truly ambiguous. We consider two heuristics to filter out suspicious repeats. First with the *size heuristic*, we assume that longer suspicious repeats are more likely to result from annotation errors than shorter ones. For instance, Table 2 displays suspicious repeats with at least 10 words that all stem from an annotation error.

Second, with the *disjoint heuristic*, we assume that actual ambiguities will be reflected in intra-corpus suspicious repeats, whereas errors will likely correspond to cases where differences in labelings are observed in different corpora. Formally, the *disjoint heuristic* flags repeats m occurring in at least two corpora A and B, and such that the set of labelings of m observed in A are disjoint from the set of labelings observed in B.

For instance, in French, *"la porte"* can either be a determiner and a noun (e.g. in the sentence *"la porte est fermée"* — the door is closed) or a pronoun followed by a verb (e.g. in the sentence *"je la porte"* — I carry her). Observing these two possible labelings in at least two corpora is a good sign of an actual ambiguity. The disjoint heuristic allows us to detect that this suspicious repeat is an actual ambiguity. To reiterate, the intuition beyond the disjoint heuristic is that for ambiguities, the two possible annotations will appear in, at least, one of the two corpora.

Conversely, systematic divergences in labeling observed across corpora are likely to be errors : for instance, in English, depending on the treebank, cardinal points are labeled as either proper nouns or as nouns. In this case, the set of labelings of the repeats in the first corpus is disjoint from the set of labeling in the second corpus and the the disjoint heuristic captures the annotation inconsistency.

Analyzing filtering heuristics To further analyze these two heuristics, we have manually annotated the suspicious repeats between the train set of the English EWT corpus and the test set of the English PUD corpus. For each suspicious repeat, we record whether it is an annotation error or an actual ambiguity. Examples of annotations are given in Table 3.

Results are in Table 4. It appears that, for the heuristics considered, a large part of the suspicious repeats correspond to annotation discrepancies rather than ambiguities. In many cases, these discrepancies result from systematic divergences in the interpretation of the UD guidelines. [5] For instance, the contraction "`n't`" is always labeled as a particle in the train set of the EWT corpus, but either as particle or an adverb in the PUD corpus. Most of these systematic differences involve distinction between nouns and proper nouns, auxiliaries and verbs and adjectives and verbs (for past participles).

4 Quantifying Annotation Divergence in the UD Corpora

4.1 Annotation Variations in the UD

We will first show how the annotation variation principle allows us to characterize the noise and/or the difficulty of PoS tagging. Table 5 reports the number of repeats and suspicious repeats in the English corpora of the UD project. These numbers have been calculated by applying the method described in the previous section to the concatenation of train, development and test sets of each treebanks. To calibrate these measures, we conducted

5. Discrepancies are not only due to improper interpretations of the guidelines, but also sometimes to actual ambiguities in the annotation rules.

heuristic	# susp. repeats	# inconsistencies
size=4	28	22 78.6%
size=3	214	153 71.5%
disjoint	580	407 70.3%
none	2507	—

Table 4: Percentage of suspicious repeats between the EWT and PUD corpora that contain an annotation inconsistency according to a human annotator either when the disjoint heuristic is used or when only suspicious repeats with at least *n* words are considered.

the same experiments with the Wall Street Journal (Marcus et al., 1993),[6] the iconic corpus of PoS tagging for which a thorough manual analysis of the annotation quality is described in (Manning, 2011).

The observations reported in Table 5 show that the number of repeats varies greatly from one corpus to another, which is not surprising considering the wide array of genres covered by the treebanks that includes sentences written by journalists or learner of English (the genres with the largest number of repeats) or sentences generated by users on social media (that contain far less repeated parts). These observations also show that the percentage of repeats that are not consistently annotated is slightly larger in the UD treebanks than in the WSJ, a corpus in which a manual inspection of the corpus reveals that many variations are 'mistakes' rather than representing uncertainties or difficulties in the PoS prediction (Manning, 2011).

More interestingly, Table 6 shows the percentage of repeats that are not consistently annotated for all possible combinations of a train and a test sets (ignoring sequences of words that do not appear at least once in both corpora). It appears that in all cases there are (sometimes significantly) more variations in annotations in cross-treebank settings than in situations where the train and the test sets belong to the same treebank. This observation suggests that there may be systematic differences in the annotations of different treebanks which could make the domain adaptation setting artificially more difficult.

Treebank	# sent.	% sent. repeat	% var.
ESL	5,124	79.0	10.4
EWT	16,622	13.1	9.0
GUM	4,399	10.5	8.5
LinES	4,564	10.9	11.8
PUD	1,000	2.7	8.7
ParTUT	2,090	18.8	9.0
WSJ	21,928	66.1	8.4

Table 5: Percentage of sentences with a repeat of at least three words in the English treebanks (`% sent. repeat`) and percentage of these repeats that are not labeled consistently (`% var.`).

6. The Penn Treebank tagset has been manually converted to the Universal PoS tagset using the mapping of (Petrov et al., 2012) generalized to the extended UD PoS tagset.

4.2 How do treebanks differ ?

To characterize the difference between two treebanks, we measure the error rate of a binary classifier deciding from which corpus an annotated sentence is coming from.[7] Intuitively, the higher this error rate, the more difficult it is to distinguish sentences of the two corpora and the more similar the treebanks are. More formally, it can be shown (Ben-David et al., 2010) that this error rate is an estimation of the $\mathscr{H}$-divergence (Kifer et al., 2004), a metric introduced in machine learning theory to quantify the impact of a change in domains by measuring the divergence between the distributions of examples sampled from two datasets.

In our experiments, we use a Naive Bayes classifier[8] and three sets of features to describe a sentence pair and their annotation : *words*, in which each example is represented by the bag of its 1-gram and 2-gram of words ; *labels*, in which examples are represented in the same way, but this time, considering PoS ; and *combi* which uses the same representation after the words of all the treebanks have been concatenated with their PoS. The first set aims at capturing a potential covariate shift, the last two target divergence in annotations. To reduce the impact of the strong between-class imbalance,[9] in all our experiments we sub-sample the largest set to ensure that the two datasets we try to distinguish always have the same number of examples. All scores in this experiment are averaged over 20 train-test splits.

7. More precisely, the classifier analyses pairs of sentences and predicts whether they belong to th same corpus or not.

8. We used the implementation provided by (Pedregosa et al., 2011) without tuning any hyper-parameters. Experiments with a logistic regression show similar results.

9. The ratio between the number of examples in the two corpora can be as large as 88.

↓ train / test →	ESL	EWT	GUM	LinES	ParTUT	WSJ	PUD
ESL	**10.0%**	11.7%	10.6%	11.1%	10.0%	12.9%	10.9%
EWT	11.8%	**8.7%**	9.1%	10.1%	8.9%	18.8%	9.2%
GUM	14.3%	9.0%	**8.2%**	11.6%	8.5%	15.8%	11.1%
LinES	16.9%	12.8%	12.6%	**12.4%**	12.5%	16.6%	14.2%
ParTUT	13.8%	10.5%	9.9%	12.0%	**9.0%**	14.9%	12.5%
WSJ	9.0%	9.9%	9.0%	9.5%	8.5%	**8.2%**	9.6%

Table 6: Percentage of repeats between a train and a test sets that are not annotated consistently. In-domain settings (i.e. when the train and test sets come from the same treebank) are reported in bold ; for each train set, the most consistent setting is underlined.

Table 7 reports the results achieved with the different features sets averaged over all combinations of a train and a test set of the same language and gives the percentage of conditions for which each feature set achieved the best results ; Figure 1 details these results for the English and French treebanks. Results for other languages show similar patterns. These results suggest that, in many cases, it is possible to accurately identify from which treebank a sentence and its annotation are coming, although these raw numbers are difficult to interpret as prediction performances are averaged over many different experimental conditions. In more than 50% of the cases, combining words to their PoS results in the best performance, which is consistent to the qualitative study reported in Section 3 : some words appear in two corpora with different PoS allowing to distinguish these corpora. This observation strongly suggests that divergence in annotations across corpora are often genuine.

5 Impact of annotation variation on prediction performance

To study annotation divergence in the UD project, we propose to analyze suspicious repeats (i.e. sequence of repeated words with different annotations). We start by extracting all the suspicious repeats that can be found when considering all the possible combinations of a train set and a test or development set of a given language. These matches are then filtered using the heuristics described in §3. There are, overall, 357,301 matches in the UD project, 69,157 of which involve 3 words or more and 14,142 5 words or more ; the disjoint heuristic selects 122,634 of these matches (see Table 8 in §A).

features	median	% best
words	78.2	31.0
labels	70.9	13.5
combi	78.8	55.5

Table 7: Precision (%) achieved over all cross-treebank conditions by a classifier identifying to which treebank a sentence belongs to.

To highlight the connection between prediction errors and annotation divergence, we compute, for each possible combination of a train and a test set (considering all languages in the UD project), the correlation between the error rate achieved on a corpus `B` when training our PoS on a corpus `A` and the number of suspicious repeats between `A` and `B` normalized by the number of tokens in `A` and `B`. The Spearman correlation coefficient between these two values is 0.72 indicating a correlation generally qualified as 'strong' following the interpretation proposed by (Cohen, 1988) : the more there are sequences of words with different annotations in the train and test sets, the worse the tagging performance, which shows that annotation inconsistencies play an important role in explaining the poor performance of PoS tagger on some conditions.

For a more precise picture, we also estimate the number of suspicious repeats that contain a prediction error. Using the disjoint heuristics to filter suspicious repeats, it appears that 70.2% (resp. 73.0%) of the suspicious repeats for English (resp. French) contain a prediction error. As expected, these numbers fall to 51.7% (resp. 49.9%) when the suspicious repeats are not filtered and therefore contain more ambiguous words. Figure 2 displays a similar trend when the suspicious repeats are filtered by their length ; similar results are observed for all other languages.

These observations suggest that annotation variations often results in prediction errors, espe-

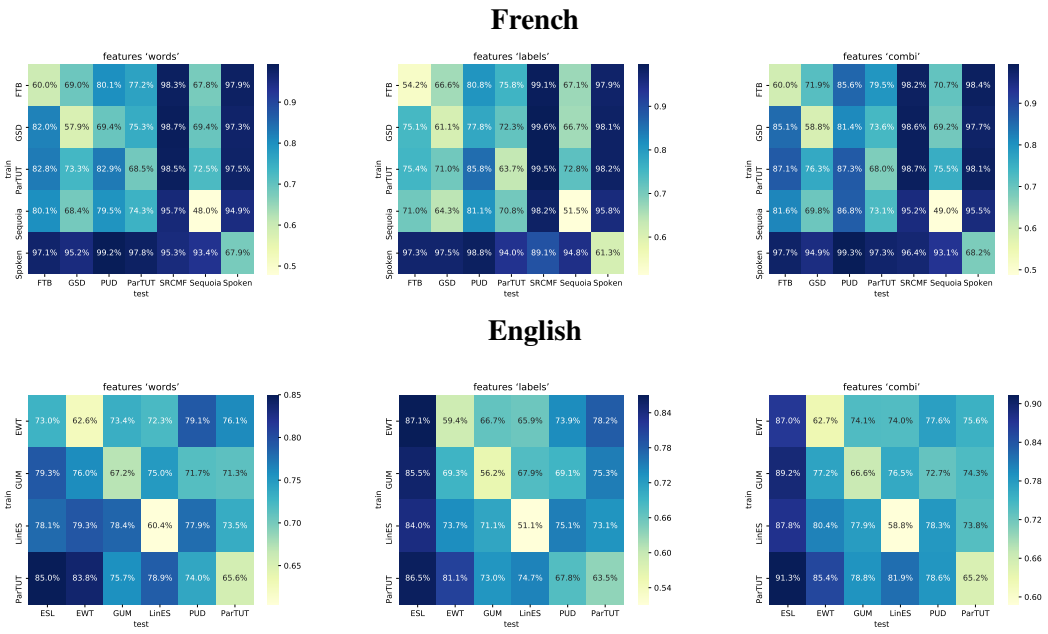

Figure 1: Precision of a classifier identifying to which French (top) or English (bottom) treebank a sentence belongs to. Train corpora are on the y-axis and test corpora on the x-axis.

cially when there are good reasons to assume that the variation actually stems from an inconsistency.

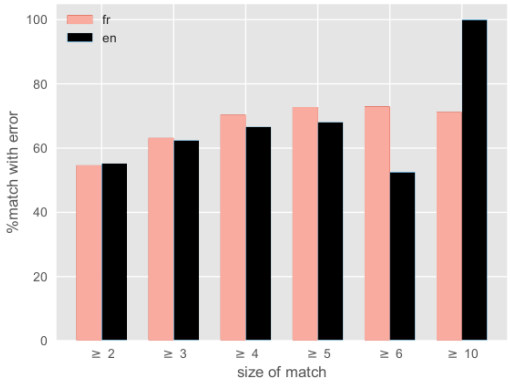

Figure 2: Percentage of suspicious repeats that contain at least one prediction error in function of their size.

6 Re-Assessing the Performance of PoS Tagger in Cross-Corpus Setting

To evaluate the impact of annotation errors on prediction performance, we propose, for each combination of a train and a test set, to train a PoS tagger and compare $\varepsilon_{\text{full}}$, the error rate achieved on the full test set to $\varepsilon_{\text{ignoring}}$ the error rate achieved ignoring errors that occur in a suspicious repeat. More precisely, $\varepsilon_{\text{ignoring}}$ is defined as :

$$\varepsilon_{\text{ignoring}} = \frac{\#\{\text{err}\} - \#\{\text{err in suspicious repeats}\}}{\#\{\text{words}\}} \quad (1)$$

where $\#\{\text{err in suspicious repeats}\}$ in the number of errors in the suspicious repeats that have survived filtering. Intuitively $\varepsilon_{\text{ignoring}}$ can be seen as an 'oracle' score corresponding to a tagger that would always predict the labels of suspicious repeat correctly. In the following, We will consider three different filters : the disjoint heuristic, keeping only suspicious repeats with more than three words and keeping all of them.

Figure 3 reports these errors rates for French and English. Results for other languages show similar results. As expected, ignoring errors in suspicious repeats significantly improve prediction performance. It even appears that $\varepsilon_{\text{ignoring}}$ is often on par with the score achieved on in-domain sets. Overall, in more than 43% (resp. 25%) of all the conditions the error rate ignoring errors in suspicious repeats filtered with the disjoint heuristic (resp. minimum heuristic) is lower than the error rate achieved on in-domain data. These values are naturally over-estimated as, in these experiments, we remove all potential annotation errors as well as words and structures that are ambiguous and therefore are more difficult to label. They can however be considered as lower-bound on the predic-

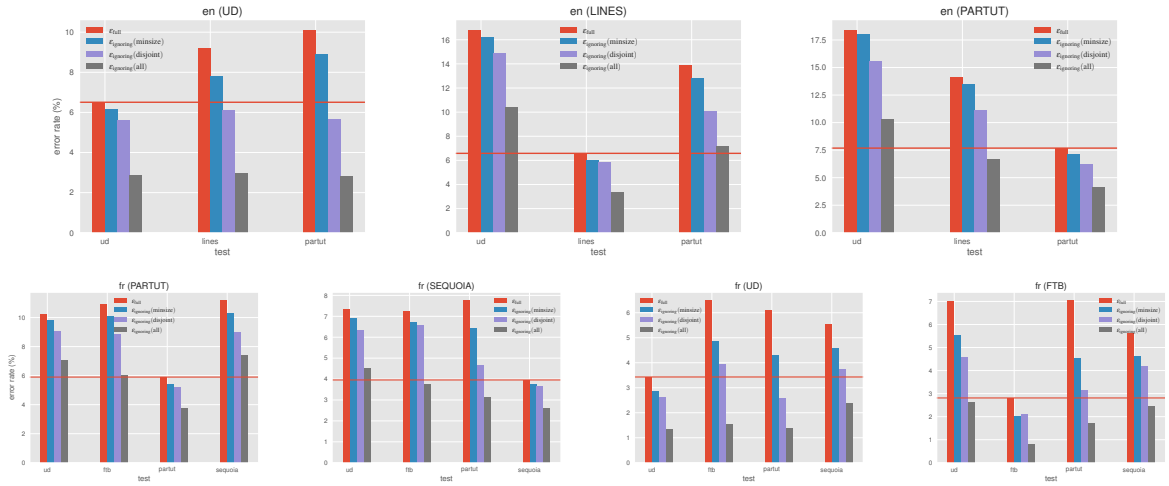

Figure 3: Error rate achieved by a PoS tagger on the different English treebanks of the UD project when errors in suspicious repeats are ignored. The red line indicates the error rate on in-domain data.

tion quality.

To assess their quality, we have manually checked all the suspicious repeats between the train set of French UD and the test set of the French FTB correcting inconsistencies and errors (almost 2,000 PoS were modified). [10] When trained on the original UD corpus, the PoS tagger achieved an error rate of 6.78% on the FTB corpus (4.51% on in-domain data). After correcting inconsistencies, the out-domain error rate falls down to 5.11%. This value is close to the error rate ignoring suspicious repeats containing three and more words, showing the validity of the heuristics we have considered.

7 Conclusion

In this work, we have shown that, for PoS tagging, many prediction errors in cross-corpora settings (which is a typical domain adaptation scenario) stem from divergence between annotations. We have also described a method to quantify this divergence. We have only considered here corpora from the UD project and PoS annotation, but we consider that our method is very generic and can be easily applied to other corpora or tasks (e.g. tokenization, dependency parsing, etc.) that we will address in future work. We also plan to see how the different experiments we have made to identify annotation errors and inconsistencies can be used during the annotation process to reduce the workload of annotators and help them creating high-quality corpora.

10. The 'corrected' corpora will be made available upon publication. In this experiment, the impact of annotation errors is under-estimated as we have only corrected errors that appear in a suspicious repeat without trying to 'generalize' these corrections to words that appear only in one corpus.

Acknowledgements

This work has been partly funded by the French *Agence Nationale de la Recherche* under ParSiTi (ANR-16-CE33-0021) and MultiSem projects (ANR-16-CE33-0013).

References

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2017. LIMSI@CoNLL'17 : UD shared task. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 163–173, Vancouver, Canada. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1-2) :151–175.

Ezra Black, Fred Jelinek, John Lafferty, David M. Magerman, Robert Mercer, and Salim Roukos. 1992. Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT'91, pages 134–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting italian treebanks : Towards an Italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Adriane Boyd, Markus Dickinson, and W. Detmar Meurers. 2008. On detecting errors in dependency

treebanks. *Research on Language and Computation*, 6(2) :113–137.

Joaquin Q. Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. *Dataset Shift in Machine Learning*. The MIT Press.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Hal Daumé III and Daniel Marcu. 2005. Learning as search optimization : Approximate large margin methods for structured prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML'05, pages 169–176, New York, NY, USA. ACM.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 107–114, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer Foster. 2010. "cba to check the spelling" : Investigating parser performance on discussion forum posts. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences : Computer Science and Computational Biology*. Cambridge University Press, New York, NY, USA.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 180–191. VLDB Endowment.

Janna Lipenkova and Milan Souček. 2014. Converting Russian dependency treebank to Stanford typed dependencies representation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2 : Short Papers*, pages 143–147, Gothenburg, Sweden. Association for Computational Linguistics.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In *Proceedings of the Computational Linguistics and Intelligent Text Processing, 12th International Conference*, CICLing 2011, pages 171–189. Springer.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english : The penn treebank. *Comput. Linguist.*, 19(2) :313–330.

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the universal dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115. Linköping University Electronic Press.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, and other. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers : a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar. Association for Computational Linguistics.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2) :227 – 244.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2017. A non-projective greedy dependency parser with bidirectional LSTMs. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 152–162, Vancouver, Canada. Association for Computational Linguistics.

Guillaume Wisniewski. 2018. Errator : a tool to help detect annotation errors in the universal dependencies project. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785, Doha, Qatar. Association for Computational Linguistics.

Yue Zhang and Joakim Nivre. 2011. Transition-based Dependency Parsing with Rich Non-local Features. In *Proceedings of ACL 2011, the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 188–193, Portland, Oregon, USA. Association for Computational Linguistics.

A Appendices

# repeated words	# repeats	# suspicious
2	4,366,885	146,516 3.36%
3	1,977,969	44,800 2.26%
4	622,192	9,684 1.56%
5	183,680	1,998 1.09%
6	60,869	509 0.84%
7	25,697	158 0.61%
8	13,132	123 0.94%
9	7,572	61 0.81%
≥ 10	24,629	264 1.07%

Table 8: Number of repeated sequence of words across the different combinations of a train set and a test set ('repeats' column) and number of these sequences that are annotated differently ('suspicious repeats' column) when no filtering is applied.