Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 1–28

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Modeling of learning curves with applications to POS tagging

Manuel Vilares Ferro *, Víctor Manuel Darriba Bilbao,
Francisco José Ribadas Pena

Department of Computer Science, University of Vigo, Campus As Lagoas s/n, 32004 Ourense, Spain

Received 16 January 2015; received in revised form 2 June 2016; accepted 3 June 2016
Available online 11 June 2016

Abstract

An algorithm to estimate the evolution of learning curves on the whole of a training data base, based on the results obtained from a portion and using a functional strategy, is introduced. We approximate iteratively the sought value at the desired time, independently of the learning technique used and once a point in the process, called prediction level, has been passed. The proposal proves to be formally correct with respect to our working hypotheses and includes a reliable proximity condition. This allows the user to fix a convergence threshold with respect to the accuracy finally achievable, which extends the concept of stopping criterion and seems to be effective even in the presence of distorting observations. Our aim is to evaluate the training effort, supporting decision making in order to reduce the need for both human and computational resources during the learning process. The proposal is of interest in at least three operational procedures. The first is the anticipation of accuracy gain, with the purpose of measuring how much work is needed to achieve a certain degree of performance. The second relates the comparison of efficiency between systems at training time, with the objective of completing this task only for the one that best suits our requirements. The prediction of accuracy is also a valuable item of information for customizing systems, since we can estimate in advance the impact of settings on both the performance and the development costs. Using the generation of part-of-speech taggers as an example application, the experimental results are consistent with our expectations.

Keywords: Correctness; Functional sequences; Learning curves; POS tagging; Proximity criterion; Robustness

1. Introduction

Creating a labeled training data base often leads to an expensive and time-consuming task, even more so when we are talking about processes involving new application domains where the resources are scarce or even non-existent. Added to this are the usually high costs of the training process itself, placing us at the origin of a bottleneck in the generation of tools based on *machine learning* (ML), as in the case of classification. The growing popularity of these techniques multiplies the need for training material, and reducing the efforts for its creation and further processing becomes a major challenge. A particularly sensitive area of work to these inconveniences is *natural language processing* (NLP), the components of which are increasingly based on ML (Biemann, 2006; Tomanek and Hahn, 2008). The problem is especially delicate on *part-of-speech* (POS) tagging, because it relies on the complexity of both the

* Corresponding author at: Department of Computer Science, University of Vigo, Campus As Lagoas s/n, 32004 Ourense, Spain. Tel.: +34 988 387280; fax +34 988 387001.

Email addresses: vilares@uvigo.es (M. Vilares Ferro), darriba@uvigo.es (V.M. Darriba Bilbao), ribadas@uvigo.es (F.J. Ribadas Pena).

http://dx.doi.org/10.1016/j.csl.2016.06.001

annotation task and the relations to be captured from learning, but also because it serves as a first step for other NLP functionalities such as parsing and semantic analysis, so errors at this stage can lower their performance (Song et al., 2012).

One way to save time and resources without loss of learning power is to anticipate the working configuration that best suits our needs, avoiding long training processes. This gives a practical meaning to our work, also providing an objective basis for defining model selection criteria (Akaike, 1973, 1974; Burnham and Anderson, 2002; Hurvich and Tsai, 1989; Lebreton et al., 1992) and model shrinking strategies (Chen et al., 2009; Sarikaya et al., 2014). Both are issues studied in practice from a statistical point of view, whose errors are expressed in probabilistic terms (Blumer et al., 1987; Floyd and Warmuth, 1995; McAllester, 1999), producing results which have proven to be quite loose (Langford, 2005) and therefore justifying the exploration of new treatment options.

To do so, we must first identify both the parameter that will serve as measure of performance for the system we are generating, generally called *accuracy*, and the factors that impact it. Unfortunately, the number and complexity of these factors is often such that it becomes impossible to weigh them. Instead, the most practical approach may be to make our own judgments in the context of the task that we are trying to accomplish (van Halteren, 1999). A way to do this is to consider the learning process as a single task to study because it compiles, relates and processes all the information provided by the user. This in turn means identifying the nature of the learning signals, in order to avoid inappropriate comparisons between ML approaches using dissimilar ones, since to a large extent it determines effectiveness. To that end, learning methods are typically classified into three categories: supervised, semi-supervised and unsupervised. The last category constitute the only possible choice when no labeled data are available, or an alternative when they are expensive to prepare, even though no useful accuracies have been achieved (Li et al., 2012). On the contrary, supervised strategies have proven to be efficient, but they require a large amount of training input that usually needs to be hand-annotated by human experts, which is an intensive task in terms of time and expertise. For its part, the semi-supervised procedures aim to combine the advantages of the previous ones when only few labeled data are available. Although the results have been mostly negative (Søgaard, 2010), the research effort is significant in the case of *active learning* (AL) (Cohn et al., 1994; Seung et al., 1992), an iterative approach that interacts with the environment in each cycle, selecting for annotation the instances which are harder to identify. Since these instances are assumed to be the most informative ones, the expected result is the acceleration of the learning process. Unfortunately, despite the potential of AL, its adoption in practice is questionable (Attenberg and Provost, 2011). In any case and having made these comments, the study of the accuracy progress during the learning process through the corresponding *learning curve* is possible, which is our starting point.

Looking ahead to POS tagging as domain to illustrate the discussion, the way to evaluate its reliability is to determine how many of the tags provided are correct, and how many superfluous ones are eliminated (van Halteren, 1999) in the case of ambiguous outputs. The absence of the latter simplifies such an evaluation, which allows us to speak of accuracy in POS tagging (DeRose, 1988). With regard to the factors that influence it, we should include any variation affecting the linguistic resources, the tag-set and the method of evaluation considered. Finally, POS tagging based on learning uses a statistical model built from labeled and/or unlabeled training data, so that we can consider unsupervised (Goldwater and Griffiths, 2007; Merialdo, 1994; Ravi and Knight, 2009), supervised (Brants, 2000; Brill, 1995; Daelemans et al., 1996; Giménez and Márquez, 2004; Schmid, 1994; Toutanova et al., 2003) and semi-supervised (Søgaard, 2010; Spoustová et al., 2009) approaches. In this latter case, the popularity of AL is growing in POS tagging tasks (Dagan and Engelson, 1995; Haertel et al., 2008; Neubig et al., 2011; Ringger et al., 2007).

Returning to the general case, one major question for both supervised and unsupervised approaches is the implementation of a condition to halt the learning process, namely to detect when it has reached its maximum or is sufficiently close for our purposes. The first goal requires a *stopping criterion* (Provost et al., 1999), while the second one implies a more general condition, referred to in this paper as *proximity criterion* and the consideration of which is a novelty in the state of the art, to the best of our knowledge. The same applies to semi-supervised strategies, although we must also add here the difficulty of designing the mechanism for selecting the instances to be annotated in every iteration when AL techniques are involved, which lies at the root of users' reluctance to adopt these methods.

All of that place the definition of stable proximity criteria at the core of our proposal. We can then talk about its *correctness* with respect to our working hypotheses when the prior estimation for accuracy is formally guaranteed in that context. We should also provide a certain capacity for assimilating the fluctuations in learning conditions without compromising the correctness, a phenomenon referred to as *variance* (Breiman, 1996b), in order to generate stable results. This is what we call the *robustness* of the model. Both issues, correctness and robustness, focus our attention in this paper, the structure of which is described below. Firstly, Section 2 examines the methodologies serving as inspiration to solve the question posed. Next, Section 3 reviews the mathematical basis necessary to support our proposal, which we introduce in Section 4. In Section 5, we describe the testing frame for the experiments illustrated in Section 6. Finally, Section 7 presents our final conclusions.

2. The state of the art

We briefly review now the main research lines in dealing with both correctness and robustness in the prediction of learning curves, highlighting their contributions and limitations, as well as their application to the NLP domain. This serves as reference for contextualizing our work.

2.1. Working on correctness

Here, the focus has been given to the convergence, leaving aside the definition of proximity criteria. Thus, current ML approaches often take for granted a set of hypotheses ensuring it, such as the access to independent and identically distributed observations (Domingo et al., 2002; Schütze et al., 2006; Tomanek and Hahn, 2008) to sample from. This allows us to assume that any learning curve is monotonic and, given that it is always bounded, the learning process has a supremum for accuracy and its convergence is guaranteed. Accordingly, the attention of researchers turns to the definition of stopping conditions in order to halt the training procedure once this value has been identified. On the basis that such functions have an initial steeply sloping portion, a more gently sloping middle portion, and a final plateau (Meek et al., 2002), the problem reduces to detect the final plateau. Initially treated from a statistical perspective (John and Langley, 1996; Valiant, 1984), this approach faces the approximation of complete learning curves as a major challenge because their complex functional forms often represent the plateau as an infinite slight incline (Frey and Fischer, 1999; Last, 2009). So, we may fit the early part of the curve well, but not the final one. Although a simple way to minimize this risk is to increase the number and size of the sample from which it is built, we should then take into account the impact in terms of computational efficiency.

For seeking a proper cost/benefit trade-off in the construction of learning curves, the researchers usually apply the principle of *maximum expected utility*, which implies expressing the problem in terms of *decision theory* (Howard, 1966). In practice, the approach differs depending upon the degree of control by the user on the process. In the absence of this control, the final cost can be defined as the sum of training data, error and model induction charges, although the interpretation may vary according to the strategy considered. Early works try to minimize induction and error costs while ignoring those of data acquisition (Provost et al., 1999), typically using a linear regression function, whose slope is compared to zero. This flaw was overcome by later proposals, first partially by assuming that the cost of cases limited the amount of training data and this amount is already specified (Weiss and Provost, 2003) to later eliminate this restriction (Weiss and Tian, 2008) and even calculate the data cost of each label feature separately through a cost-sensitive learner (Sheng and Ling, 2007). Nonetheless, finding the global optimum of total cost can in no case be guaranteed (Last, 2009), and it is not infrequent that one wants to stop only when the desired degree of accuracy is met, which would bring us back to the consideration of pure stopping criteria.

A separate case is that of AL, where we select the training examples to label and even quantify costs associated to specific feature values, which lead us to take into account the cost of teacher and the cost of tests (Turney, 2000), respectively. However, since much of the work on AL assumes a fixed budget (Kapoor and Greiner, 2005), practical stopping criteria often rely simply on measuring the confidence of the learning process. Given that most of the research focuses on *pool-based active learning* (Lewis and Gale, 1994), in which the selection is made from a pool, such a measure applies either to a separate data set (Vlachos, 2008) or to the unlabeled data pool (Zhu and Ma, 2012). The assumption in the first case is that we cannot take advantage of the remaining instances in the pool when that confidence drops, while in the second one the starting point is the uncertainty of the classifier (Roy and McCallum, 2001). The point where the pool becomes uninformative can also be determined through the gradient of the performance (Laws and Schütze, 2008), whose rising estimation slows to an almost horizontal slope at about the time when the learning reaches its peak. We then stop the process when that gradient approaches zero.

2.2. Working on robustness

Turning now to robustness, proposals pass through the generation of different versions (weak predictors) of the learning curve by changing the distribution of the training repeatedly, which makes it possible to combine by aggregation the set of hypotheses so generated. In the case of *bagging*[1] procedures (Breiman, 1996a), the weak predictors are built in parallel and then combined using voting (classification) (Leung and Parker, 2003) or averaging (regression) (Leite and Brazdil, 2007). On the contrary, *boosting* algorithms (Schapire, 1990) do it sequentially, which allows to adapt the distribution of the training data base from the performance of the previous weak predictors. This gives rise to *arcing*[2] strategies (Freund and Schapire, 1996), where increasing weight is placed on the more frequently misclassified observations. Since these are the troublesome points, focusing on them may do better than the neutral bagging approach (Bauer and Kohavi, 1999), justifying its popularity (García-Pedrajas and De Haro-García, 2014). Another common strategy for increasing stability, especially in AL, is the use of thresholds providing a flexibility with regard to variance, but without warranty of any type. The stability of predictions is studied on a set of examples, called the stop set, that do not have to be labeled (Bloodgood and Vijay-Shanker, 2009). At best, the dynamic update of thresholds is outlined to increase soundness across changing data sets (Zhu and Ma, 2012).

2.3. An overview for the NLP domain

The use of learning curves for anticipating the performance of NLP tools has been the subject of ongoing research during the last years, mainly in the sphere of *machine translation* (MT). So, they have been employed for assessing the quality of MT systems (Bertoldi et al., 2012; Turchi et al., 2008), for optimizing parameter setting (Koehn et al., 2003) and for estimating how many training data are required to achieve a certain degree of translation accuracy (Kolachina et al., 2012), although no formal stopping criteria are described. They were also used to evaluate the impact of a concrete set of distortion factors on the performance of a concrete operational model (Birch et al., 2008), albeit with meager results. Together, all these works recall the essence of our problem formulation and serve also as examples of how correctness and robustness are treated in the prediction of learning curves. So, the lack of well-founded mathematical models results in proposals whose sole support comes from a test battery, which in the best case offers a partial vision of the problem. Therefore, they are a long way from achieving a solution verifying our requirements on correctness.

In the case of AL techniques, their popularity is growing in POS tagging tasks (Dagan and Engelson, 1995; Haertel et al., 2008; Ringger et al., 2007) and closely related ones such as named entity recognition (Laws and Schütze, 2008; Shen et al., 2004; Tomanek et al., 2007) or word sense disambiguation (Chan and Ng, 2007; Chen et al., 2006; Zhu and Hovy, 2007), with the purpose of reducing the annotation effort. The same is true for information extraction (Culotta and McCallum, 2005; Thompson et al., 1999), parsing (Becker and Osborne, 2005; Tang et al., 2002) or text classification (Lewis and Gale, 1994; Liere and Tadepalli, 1997; McCallum and Nigam, 1998; Tong and Koller, 2002) applications. In none of these cases, however, these works have contributed to the treatment of correctness and robustness beyond what we have seen so far.

2.4. Our contribution

In order to confer reliability on the prediction of accuracy in tools resulting from ML processes, we introduce an iterative functional architecture as an alternative to the classic statistical techniques. This is defined on a sequence of approximations for the partial learning curves, which are calculated from an increasing set of observations. On the basis of a set of working hypotheses widely recognized in both learning curves and training data bases, the correctness of the method is theoretically proven. We can then, in contrast to earlier works, define a proximity criterion to stop the learning once a degree of accuracy fixed by the user is reached. Regarding robustness against variations in the working hypotheses, we propose an anchoring mechanism which formally limits their impact, and is fully compatible with the basic algorithm.

[1] For *bootstrap aggregating*.

[2] For *adaptive resampling and combining*.

3. The formal framework

The aim now is to describe our abstract model on a mathematical basis that would enable us to prove its correctness. We choose a function, which must be continuous so that it provides sustainability to estimate in advance the learning curve for accuracy. Since the approximations of that function can be modeled from partial learning curves, it seems natural to raise its calculation as the convergence of the sequence of such approximations while the training process advances.

3.1. The mathematical support

We first recall some notions by Apostol (2000) on the theory of sequences in the real metric space $(\mathbb{R}, |\ |)$, where $|\ |$ denotes the Euclidean distance defined as the absolute difference, and $\mathbb{R}$ is the set of real numbers. For the sake of simplicity, we assume familiarity with the concepts of continuity and derivability of a real function. We denote the set of natural numbers by $\mathbb{N}$, and we assume that $0 \notin \mathbb{N}$.

Definition 1 *Let* $\{x_i\}_{i \in \mathbb{N}}$ *be a sequence in* $(\mathbb{R}, |\ |)$, *we say that it is a sequence convergent to* $x_\infty \in \mathbb{R}$ *iff*

$$\forall \varepsilon > 0, \quad \exists n \in \mathbb{N}, \quad \forall i \geq n \Rightarrow |x_i - x_\infty| < \varepsilon \tag{1}$$

where x_∞ *is called the* limit of $\{x_i\}_{i \in \mathbb{N}}$, *using the notation* $\lim\limits_{i \to \infty} x_i = x_\infty$.

A sequence converges when we can situate, from a given moment, all its elements as close to the limit as we want to. Furthermore, it has been proven that any monotonic increase (resp. decrease) and upper (resp. lower) bounded sequence converges to its supremum (resp. infimum). Since we want to study the convergence of a collection of curves, we need to extend the concept to sequences of real functions.

Definition 2 *Let* $\Delta := \{f : E \subseteq \mathbb{R} \to \mathbb{R}\}$ *and let* $\{f_i\}_{i \in \mathbb{N}}, f_i \in \Delta$ *be a sequence of functions, we say that it is a sequence pointwise convergent to* $\mathrm{f}_\infty \in \Delta$ *iff*

$$\forall x \in E, \quad \varepsilon > 0, \quad \exists n_x \in \mathbb{N}, \quad \forall i \geq n_x \Rightarrow |f_i(x) - f_\infty(x)| < \varepsilon \tag{2}$$

where f_∞ *is called the* pointwise limit of $\{f_i\}_{i \in \mathbb{N}}$, *using the notation* $\lim\limits_{i \to \infty}{}^{p} f_i = f_\infty$.

A sequence of functions is pointwise convergent if the sequence of their values on each instant converges, which implies that we can calculate the limit point-to-point, although the speed of convergence may be different in each case. This poses a problem when we want to use the limit function for prediction purposes because the results obtained could vary greatly even over points close to the observations. Namely, in order to provide reliability to our estimates, we need a criterion making it possible to consider a common convergence threshold for the whole functional domain considered.

Definition 3 *Let* $\Delta := \{f : E \subseteq \mathbb{R} \to \mathbb{R}\}$ *and let* $\{f_i\}_{i \in \mathbb{N}}, f_i \in \Delta$ *be a sequence of functions, we say that it is a* sequence uniformly convergent to $\mathrm{f}_\infty \in \Delta$ *iff*

$$\forall \varepsilon > 0, \quad \exists n \in \mathbb{N}, \quad \forall i \geq n \Rightarrow |f_i(x) - f_\infty(x)| < \varepsilon, \quad \forall x \in E \tag{3}$$

where f_∞ *is called the* uniform limit of $\{f_i\}_{i \in \mathbb{N}}$, *using the notation* $\lim\limits_{i \to \infty}{}^{u} f_i = f_\infty$.

This identifies functional sequences for which all points converge at the same pace to the limit, and allows its continuity to be inferred when the curves in the sequence are continuous. Such a property is also useful for prediction purposes since continuity is a guarantee of regularity, matching small variations in the input to small ones in the output.

3.2. The working hypotheses

Since the approximation of partial learning curves is the basis for our proposal, we need to gather as much information as possible about their nature in order to identify the functions better adapted to that requirement. Thus, the

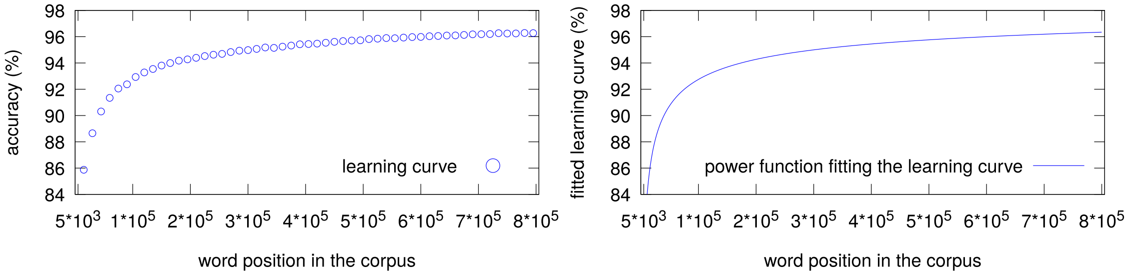

Fig. 1. Learning curve for the training process of fnTBL on Frown corpus, and an accuracy pattern fitting it.

point of departure for generating a learning curve is a sequence of observations calculated from a series of cases incrementally taken from a training data base over the time, which should meet certain conditions in order to make the prediction task possible. Basically, we assume this data base to be independently and identically distributed (Schütze et al., 2006; Tomanek and Hahn, 2008), in order to obtain a predictable progression of the estimation trace for accuracy over a virtually infinite interval. This does not imply a loss of generality since we can re-order the training data base (Clark et al., 2010) and generate as many observations as we want.

We then accept that a learning curve is a positive definite and strictly increasing function on $\mathbb{N}$, where natural numbers represent the sample size, upper bounded by 100. We also assume that the speed of increase is higher in its first stretch, where the learning is faster, decreasing as the training process advances and giving the curve a concave form with a horizontal asymptote (Frey and Fischer, 1999; Last, 2009; Meek et al., 2002). One might then argue that such requirements cannot be completely guaranteed in practice, as it can be observed in the left-most diagram of Fig. 1, which shows the learning curve for the training process of the *fast transformation-based learning* (fnTBL) tagger (Ngai and Florian, 2001) on the *Freiburg-Brown* (Frown) corpus of American English (Mair and Leech, 2007). We consider here examples indicated by the position of a word from the beginning of the text,[3] thus delimiting the section of this used to generate them and evidencing the existence of small irregularities in both concavity and increase. It is then necessary to take into account that the idealization is inherent to the scientific method, the objective of which is to lay the foundations for the correctness, leaving for a subsequent stage the question of robustness.

3.3. The notational support

Having identified the context of the problem, we need to formally capture the data structures we are going to work with, such as the collection of instances on which the accuracy measurements are sequentially applied to generate the observations that will serve as starting point for the prediction task.

Definition 4 *Let $\mathcal{D}$ be a training data base, $\mathcal{K} \subsetneq \mathcal{D}$ a non empty set of initial items and $\sigma : \mathbb{N} \to \mathbb{N}$ a function. We define a learning scheme for $\mathcal{D}$ with kernel $\mathcal{K}$ and step function σ, as a triple $\mathcal{D}_\sigma^\mathcal{K} = [\mathcal{K}, \sigma, \{\mathcal{D}_i\}_{i \in \mathbb{N}}]$ with $\{\mathcal{D}_i\}_{i \in \mathbb{N}}$ an incremental cover of $\mathcal{D}$ verifying:*

$$\mathcal{D}_1 := \mathcal{K} \quad \textit{and} \quad \mathcal{D}_i := \mathcal{D}_{i-1} \cup \mathcal{I}_i, \quad \mathcal{I}_i \subset \mathcal{D} \setminus \mathcal{D}_{i-1}, \quad \|\mathcal{I}_i\| = \sigma(i), \quad \forall i \geq 2 \tag{4}$$

where $\|\mathcal{I}_i\|$ is the cardinality of $\mathcal{I}_i$. We refer to $\mathcal{D}_i$ as the individual of level i for $\mathcal{D}_\sigma^\mathcal{K}$.

A learning scheme relates a level $i \in \mathbb{N}$ with the position $x_i := \|\mathcal{D}_i\|$ in the training data base of its corresponding case, thereby determining the sequence of observations, $\{[x_i, \mathcal{A}_{\otimes\otimes\otimes}[\mathcal{D}_\sigma^\mathcal{K}](x_i)], x_i := \|\mathcal{D}_i\|\}_{i \in \mathbb{N}}$ where $\mathcal{A}_{\otimes\otimes\otimes}[\mathcal{D}_\sigma^\mathcal{K}](x_i)$ is the accuracy achieved on such instance by the system we are studying. Namely $\mathcal{A}_{\otimes\otimes\otimes}[\mathcal{D}_\sigma^\mathcal{K}]$ is the learning curve associated to the scheme $\mathcal{D}_\sigma^\mathcal{K}$ that we want to approximate, and the kernel $\mathcal{K}$ delimits a portion of $\mathcal{D}$ we believe to be enough to initiate consistent evaluations. We therefore need a functional pattern serving as model for fitting these curves. This leads us to consider candidates verifying our working hypotheses, but also representing real C-infinity functions in order to provide reliability to our estimates through their structural smoothness.

[3] All the word counts in this paper include punctuation marks.

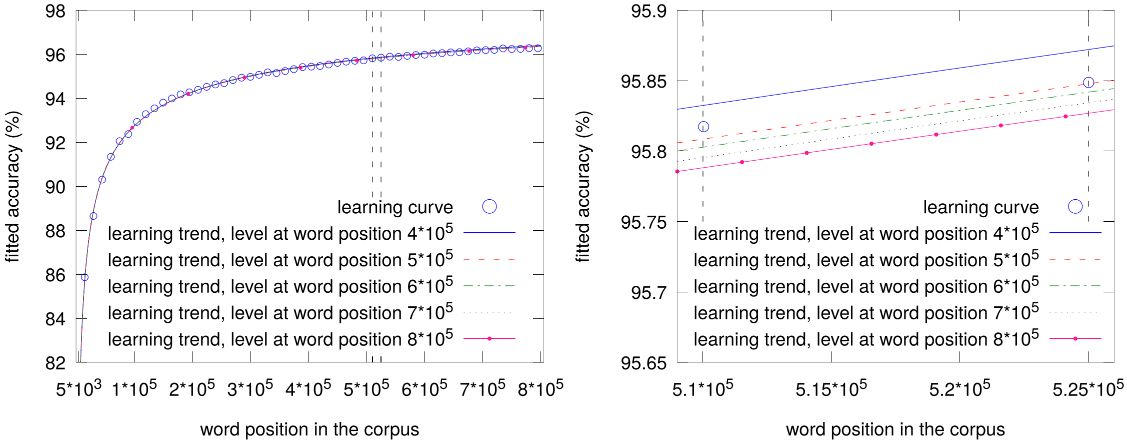

Fig. 2. Learning trace for the training process of fnTBL on Frown, with details in zoom.

Definition 5 *Let $C^{\infty}_{(0,\infty)}$ be the real C-infinity functions in $\mathbb{R}^{+}$, we say that $\pi : \mathbb{R}^{+^n} \rightarrow C^{\infty}_{(0,\infty)}$ is an accuracy pattern iff $\pi(a_1, \ldots, a_n)$ is positive definite, concave and strictly increasing.*

An example of accuracy pattern is the *power family* of curves $\pi(a,b,c)(x) := -a * x^{-b} + c$, hereinafter used as running example. They have $\lim_{x\rightarrow\infty} \pi(a,b,c)(x) = c$ as horizontal asymptote and verify:

$$\pi(a,b,c)'(x) = a * b * x^{-(b+1)} > 0 \quad \pi(a,b,c)''(x) = -a * b * (b+1) * x^{-(b+2)} < 0 \tag{5}$$

which guarantees increase and concavity in $\mathbb{R}^{+}$, respectively. This is illustrated in the right-most curve of Fig. 1, whose goal is fitting the learning curve represented in the left-hand side. Here, the values $a = 542.5451$, $b = 0.3838$ and $c = 99.2876$ are provided by the *trust region method* (Branch et al., 1999), a regression technique minimizing the summed square of *residuals*, namely the differences between the observed values and the fitted ones.

Definition 6 *Let $\mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ be a learning curve, π an accuracy pattern and $\ell \in \mathbb{N}, \ell \geq 3$ an item position in the training data base $\mathcal{D}$. We define the learning trend of level ℓ for $\mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ using π, as a curve $\mathcal{A}^{\pi}_{\ell}[\mathcal{D}^{\mathcal{K}}_{\sigma}] \in \pi$, fitting the observations $\{[x_i, \mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x_i)], x_i := \|\mathcal{D}_i\|\}^{\ell}_{i=1}$. A sequence of learning trends $\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}] := \{\mathcal{A}^{\pi}_{\ell}[\mathcal{D}^{\mathcal{K}}_{\sigma}]\}_{\ell\in\mathbb{N}}$ is called a learning trace. We denote by $\rho_{\ell}(i) := [\mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}] - \mathcal{A}^{\pi}_{\ell}[\mathcal{D}^{\mathcal{K}}_{\sigma}]](\|\mathcal{D}_i\|)$ the residual of $\mathcal{A}^{\pi}_{\ell}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ at the level $i \in \mathbb{N}$. We refer to $\{\alpha_{\ell}\}_{\ell\in\mathbb{N}}$ as the asymptotic backbone of $\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$, where $y = \alpha_{\ell}$ is the asymptote of $\mathcal{A}^{\pi}_{\ell}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$.*

The learning trends also fit, namely approximate, the partial learning curves beyond their own level. However, this is not enough to provide a credible prediction for accuracy because to do so requires information on their evolution during the training process. Learning traces solve this question, allowing the extraction of a sequence of fitted values for a case, representing the evolution of the estimation at that instant. Continuing with the example for the tagger fnTBL and the corpus Frown, Fig. 2 shows a portion of the learning trace associated to a kernel and constant step function $5*10^3$, the levels of which are indicated by the corresponding word position in the corpus, including also a more detailed view. Finally, we can only consider learning trends from the level $\ell = 3$ because we need at least three observations to generate a curve.

4. The abstract model

We lay the theoretical foundations of our proposal to later interpret them from an operational point of view. The first step is proving its correctness, which also allows us to discuss it against the deviations introduced by real observations.

4.1. Correctness

Since the intention is for us to reliably approximate the learning curve $\mathcal{A}_{\infty\!\!\infty}[\mathcal{D}_\sigma^\mathcal{K}]$ by means of the limit function of the sequence $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}] := \{\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]\}_{i \in \mathbb{N}}$ of learning trends incrementally built from the observations, we start by studying its uniform convergence. As notational facility, we extend the natural order in $\mathbb{N}$ in such a way that $\infty\!\!\infty > \infty > i > 0, \forall i \in \mathbb{N}$

Theorem 1 *Let* $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ *be a learning trace, with or without anchors. Then its asymptotic backbone is monotonic and* $\mathcal{A}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}] := \lim_{i \to \infty}^{u} \mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ *exists, is positive definite, increasing, continuous and upper bounded by 100 in* $(0,\infty)$.

Proof: Having fixed a level $i \in \mathbb{N}$, the fitting algorithm minimizes a weighting function on the set of residuals $\{\rho_i(j)\}_{j \le i}$ in order to generate a learning trend $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$, such that $\sum_{j \le i} \rho_i(j) = 0$. Consequently, the latter intersects necessarily the learning curve $\mathcal{A}_{\infty\!\!\infty}[\mathcal{D}_\sigma^\mathcal{K}]$. Since the concavity of both does not vary in $(0,\infty)$, they cross at one ($q_{\infty\!\!\infty}^i$) or two ($p_{\infty\!\!\infty}^i$ and $q_{\infty\!\!\infty}^i$) points in that interval. As shown in the left-hand-side of Figs. 3 and 4, this delimits two ($B_{\infty\!\!\infty}^i$ and $C_{\infty\!\!\infty}^i$) or three ($A_{\infty\!\!\infty}^i$, $B_{\infty\!\!\infty}^i$ and $C_{\infty\!\!\infty}^i$) sub-intervals. So, $A_{\infty\!\!\infty}^i := (0, p_{\infty\!\!\infty,x}^i]$, $B_{\infty\!\!\infty}^i := (p_{\infty\!\!\infty,x}^i, q_{\infty\!\!\infty,x}^i]$ and $C_{\infty\!\!\infty}^i := (q_{\infty\!\!\infty,x}^i, \infty)$, where $p_{\infty\!\!\infty}^i := (p_{\infty\!\!\infty,x}^i, p_{\infty\!\!\infty,y}^i)$, $q_{\infty\!\!\infty}^i := (q_{\infty\!\!\infty,x}^i, q_{\infty\!\!\infty,y}^i)$, and the sign of residuals in $B_{\infty\!\!\infty}^i$ is different from that of $A_{\infty\!\!\infty}^i$ and $C_{\infty\!\!\infty}^i$. Thus, the trend in $C_{\infty\!\!\infty}^i$ is either above or below the observation and $0 < p_{\infty\!\!\infty,x}^i < q_{\infty\!\!\infty,x}^i < \|\mathcal{D}_i\|$

As learning curves and trends are strictly monotonic (increasing), the impact of the observations in the fitting process always applies in one direction from the first position $q_{\infty\!\!\infty,x}^1$, increasing as the levels ascend. Accordingly, the asymptotic backbone $\alpha := \{\alpha_i\}_{i \in \mathbb{N}}$ is also monotonic. More specifically, as illustrated in Fig. 3 (resp. Fig. 4), α is lower (resp. upper) bounded by the asymptotic value of the learning curve $\alpha_{\infty\!\!\infty}$ if the sequence is decreasing (resp. increasing). Given that this bound is, in fact, an infimum (resp. a supremum), we can thus conclude that α converges monotonically to $\alpha_{\infty\!\!\infty}$.

On the other hand, $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ is itself a learning trend approximating $\mathcal{A}_j^\pi[\mathcal{D}_\sigma^\mathcal{K}], i < j$. Following the same reasoning used before with regard to the learning curve, they cross at one (q_j^i) or two (p_j^i and q_j^i) points in $(0,\infty)$. This delimits two (B_j^i and C_j^i) or three (A_j^i, B_j^i and C_j^i) sub-intervals. So, $A_j^i := (0, p_{j,x}^i]$, $B_j^i := (p_{j,x}^i, q_{j,x}^i]$ and $C_j^i := (q_{j,x}^i, \infty)$, where $p_j^i := (p_{j,x}^i, p_{j,y}^i)$, $q_j^i := (q_{j,x}^i, q_{j,y}^i)$, and $0 < p_{j,x}^i < q_{j,x}^i < \|\mathcal{D}_j\|$.

Let $\{(q_{i,x}^{i-1}, q_{i,y}^{i-1})\}_{i \ge 4}$ (resp. $\{(p_{i,x}^{i-1}, p_{i,y}^{i-1})\}_{i \ge 4}$) then be the sequence of the last (resp. first, if they exist) points of intersection between a trend $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ and the previous one $\mathcal{A}_{i-1}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$. Given that α converges monotonically, $\{q_{i,y}^{i-1}\}_{i \ge 4}$ also does so, as shown in Figs. 3 and 4 regardless of the type of monotony exhibited by the asymptotic backbone. For the same reason, if it exists, $\{p_{i,y}^{i-1}\}_{i \ge 4}$ converges to a point less than or equal to the one for $\{q_{i,y}^{i-1}\}_{i \ge 4}$. Since $\{p_{i,y}^{i-1}\}_{i \ge 4}$ and $\{q_{i,y}^{i-1}\}_{i \ge 4}$ are monotonic, $\{p_{i,x}^{i-1}\}_{i \ge 4}$ and $\{q_{i,x}^{i-1}\}_{i \ge 4}$ are also monotonic because the trends are strictly increasing. Moreover,

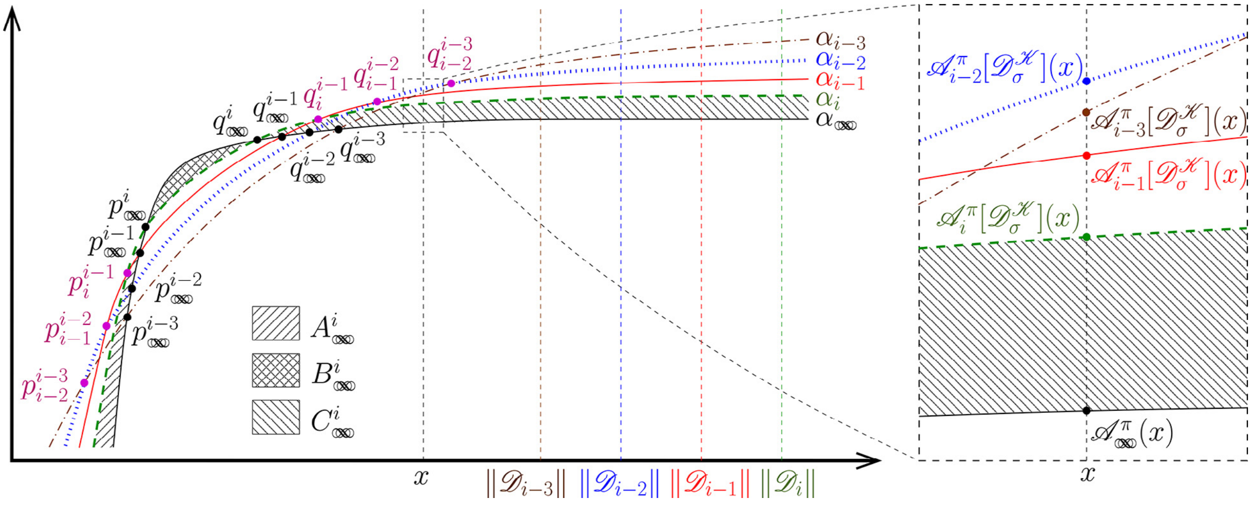

Fig. 3. An example of construction for a decreasing learning trace, with details in zoom.

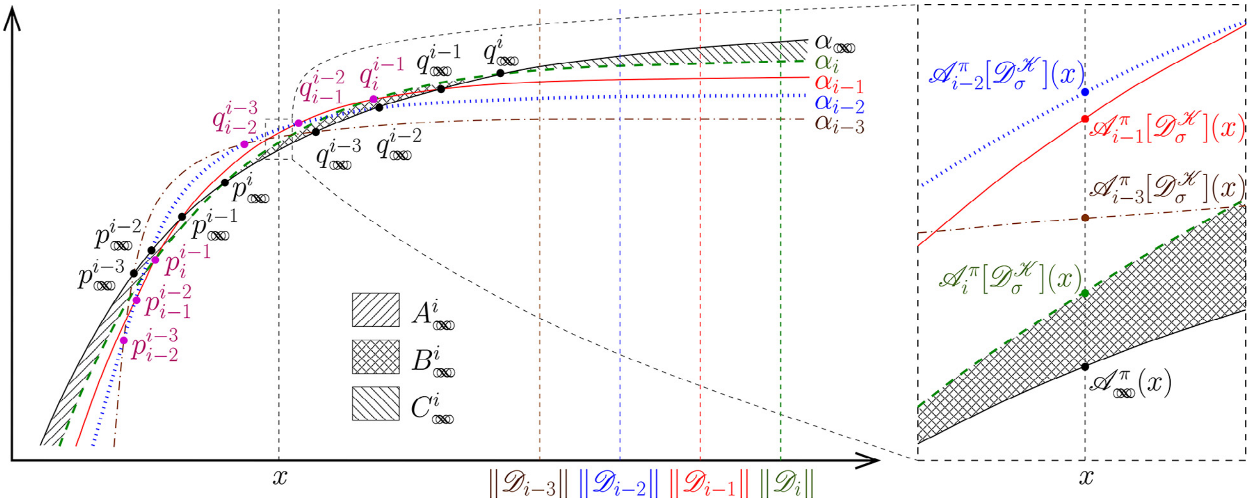

Fig. 4. An example of construction for an increasing learning trace, with details in zoom.

$0 < p^{i-1}_{i,x} < q^{j-1}_{j,x} < \|\mathcal{D}_j\|, \forall i, j \in \mathbb{N}, i \neq j$ because otherwise $\{q^{i-1}_{i,y}\}_{i \geq 4}$ would not be monotonic. This implies that $\{p^{i-1}_{i,x}\}_{i \geq 4}$ is also lower (resp. upper) bounded by 0 (resp. $q^{1}_{2,x}$) and, therefore, $\lim\limits_{i \to \infty} p^{i-1}_{i,x}$ exists.

Thus, as illustrated in the right-hand-side of Figs. 3 and 4, in $\left(0, \lim\limits_{j \to \infty} p^{j-1}_{j,x}\right)$ the sequences $\{\mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)\}_{i \in \mathbb{N}}$ are monotonic from i such that $x < p^{i-1}_{i,x}$; in $\left[\lim\limits_{j \to \infty} p^{j-1}_{j,x}, \lim\limits_{j \to \infty} q^{j-1}_{j,x}\right)$ from i such that $x < q^{i-1}_{i,x}$; and in $\left[\lim\limits_{j \to \infty} q^{j-1}_{j,x}, \infty\right)$ if $\{q^{j-1}_{j,x}\}_{j \geq 4}$ converges. Since they are also bounded because the learning trends are particular configurations of π, $\mathcal{A}^{\pi}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x) := \lim\limits_{i \to \infty} \mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)$ is a well defined, positive definite and increasing function. Trivially, $\mathcal{A}^{\pi}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}] = \lim\limits_{i \to \infty}{}^{p} \mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ in $(0,\infty)$. Given that $\{\alpha_i\}_{i \in \mathbb{N}}$ converges to $\alpha_{\infty\!\!\infty}$ and the observations are upper bounded by 100, we have

$$\exists n_{\alpha} \in \mathbb{N}, \quad \forall i \geq n_{\alpha} \Rightarrow \left|\mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x) - \alpha_{\infty\!\!\infty}\right| \leq \varepsilon = |100 - \alpha_{\infty\!\!\infty}|, \quad \forall x \in (0, \infty) \tag{6}$$

or, in other words,

$$\exists n_{\alpha} \in \mathbb{N}, \quad \forall i \geq n_{\alpha} \Rightarrow \left|\mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)\right| \leq |\alpha_i| \leq 100, \quad \forall x \in (0, \infty) \tag{7}$$

which implies that $\mathcal{A}^{\pi}_{\infty}[\mathcal{D}\mathcal{K}_{\sigma}]$ is also upper bounded by 100.

To prove now that the convergence is uniform, it is sufficient to take into account that

$$\lim_{i \to \infty}\left(\sup_{x \in (0,\infty)} \left|\mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x) - \mathcal{A}^{\pi}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)\right|\right) = 0 \tag{8}$$

because, by construction, no vertical asymptotic behavior is observable in the trace. As all the trends are continuous, then so is their uniform limit $\mathcal{A}^{\pi}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$. ■

All this provides us with an abstract model to estimate the learning curve $\mathcal{A}_{\infty\!\!\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ over the training data base by iteratively approximating the function $\mathcal{A}^{\pi}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$, while there is a need for measuring the convergence (resp. error) threshold at each stage of the process, in order to give it a practical sense. Namely, after fixing a level $i \in \mathbb{N}$, we have to calculate an upper bound for the distance between $\mathcal{A}^{\pi}_{j}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ and $\mathcal{A}^{\pi}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ (resp. $\mathcal{A}_{\infty\!\!\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$) in the interval $[\|\mathcal{D}_j\|, \infty), \forall j \geq i$. In other words, we need to define a proximity criterion.

Theorem 2 (Correctness) *Let* $\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ *be a learning trace, with* $\mathrm{y} = \alpha_{\mathrm{i}}$ *the asymptote for* $\mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}], \quad \forall i \in \mathbb{N} \cup \{\infty, \infty\!\!\infty\}$ *and* $(q^{i-1}_{i,x}, q^{i-1}_{i,y})$ *the last point in* $\mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}] \cap \mathcal{A}^{\pi}_{i-1}[\mathcal{D}^{\mathcal{K}}_{\sigma}], \quad \forall i \geq 4$. *We then have*

$$\left|\left[\mathcal{A}_k^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]-\mathcal{A}_j^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right](x)\right| \leq \varepsilon_i := \left|q_{i,y}^{i-1}-\alpha_i\right|, \quad \forall k,\, j \geq i \geq 4, \quad \forall x \in \left[q_{i,x}^{i-1}, \infty\right) \tag{9}$$

$$\left(\text{resp.} \left|\left[\mathcal{A}_k^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]-\mathcal{A}_j^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right](x)\right| \leq \varepsilon_i := \left|q_{\infty\infty,y}^{i}-\alpha_{\infty\infty}\right|, \forall k,\, j \geq i \geq 1, \forall x \in \left[q_{\infty\infty,x}^{i}, \infty\right)\right) \tag{10}$$

if $\{\alpha_i\}_{i\in\mathbb{N}}$ *is decreasing* (resp. increasing)*, with* $\{\varepsilon_i\}_{i\in\mathbb{N}\cup\{\infty,\infty\infty\}}$ *monotonically decreasing and convergent to* 0.

Proof: Let us first suppose $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing. As shown in the left-hand-side of Fig. 3, we follow from the construction of $\mathcal{A}_\infty^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]$ that

$$q_{i,y}^{i-1} := \mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\left(q_{i,x}^{i-1}\right) \leq \mathcal{A}_l^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right](x) \leq \alpha_i, \quad \forall l \geq i \geq 4, \quad \forall x \in \left[q_{i,x}^{i-1}, \infty\right) \tag{11}$$

from which Equation 9 is trivial. Following Theorem 1, as $\{q_{i,y}^{i-1}\}_{i\geq 4}$ is increasing and $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing with $q_{i,y}^{i-1} < \alpha_i$, the sequence $\{\varepsilon_i\}_{i\in\mathbb{N}\cup\{\infty,\infty\infty\}}$ is decreasing. Since $\lim\limits_{i\to\infty}\varepsilon_i := \lim\limits_{i\to\infty} q_{i,y}^{i-1} - \lim\limits_{i\to\infty}\alpha_i = \alpha_{\infty\infty} - \alpha_{\infty\infty} = 0$, the thesis is then proven.

In an analogous manner, as shown in the left-hand-side of Fig. 4 and when $\{\alpha_i\}_{i\in\mathbb{N}}$ is increasing, it verifies

$$q_{\infty\infty,y}^{i} := \mathcal{A}_{\infty\infty}\left[\mathcal{D}_\sigma^\mathcal{K}\right]\left(q_{\infty\infty,x}^{i}\right) \leq \mathcal{A}_l^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right](x) \leq \alpha_{\infty\infty}, \quad \forall l \geq i \geq 1, \quad \forall x \in \left[q_{\infty\infty,x}^{i}, \infty\right) \tag{12}$$

from which Equation 10 is trivial. Following a similar reasoning to the one applied to $\{q_{i,y}^{i-1}\}_{i\geq 4}$ in Theorem 1, we deduce that $\{q_{\infty\infty,y}^{i}\}_{i\geq 1}$ is monotonic (decreasing), so $\{\varepsilon_i\}_{i\in\mathbb{N}\cup\{\infty,\infty\infty\}}$ is monotonically decreasing. Finally, $\lim\limits_{i\to\infty}\varepsilon_i := \lim\limits_{i\to\infty}\mathcal{A}_{\infty\infty}\left[\mathcal{D}_\sigma^\mathcal{K}\right]\left(q_{\infty\infty,y}^{i}\right) - \alpha_{\infty\infty} = \lim\limits_{x\to\infty}\mathcal{A}_{\infty\infty}\left[\mathcal{D}_\sigma^\mathcal{K}\right](x) - \alpha_{\infty\infty} := \alpha_{\infty\infty} - \alpha_{\infty\infty} = 0$. ■

Given that $q_{i,x}^{j} < \min\{\|\mathcal{D}_i\|, \|\mathcal{D}_j\|\}, \quad \forall i, j \in \mathbb{N}\cup\{\infty, \infty\infty\}, i \neq j$, this result meets our requirements. Unfortunately, it has only a practical reading in the first case, when the convergence is decreasing. Otherwise, the upper bound depends on $\alpha_{\infty\infty}$, which is the final value for accuracy we want to estimate and thus is unknown. In order to break the deadlock, we have no option but to find a criterion for correctness based on the individual behavior of each learning trend as part of the approximation process.

Definition 7 *Let* $\mathcal{A}^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]$ *be a learning trace with asymptotic backbone* $\{\alpha_i\}_{i\in\mathbb{N}}$*. We define the layer of convergence for* $\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right], i \in \mathbb{N}$ *as the value* $\chi\left(\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right) := \left|\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right](\|\mathcal{D}_i\|) - \alpha_i\right|$.

This concept allows us to measure the contribution of each learning trend to the convergence process, which provides the key for a practical interpretation of the correctness, regardless of the type of monotony associated to the asymptotic backbone.

Theorem 3 (Layered Correctness) *Let* $\mathcal{A}^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]$ *be a learning trace with asymptotic backbone* $\{\alpha_i\}_{i\in\mathbb{N}}$*. We then have*

$$\forall \varepsilon > 0, \quad \exists n \in \mathbb{N}, \quad \textit{such that} \quad \left[\chi\left(\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right) \leq \varepsilon \Leftrightarrow i \geq n\right] \tag{13}$$

Proof: Following the proof of Theorem 1, both sequences $\{\alpha_i\}_{i\in\mathbb{N}}$ and $\{\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right](\|\mathcal{D}_i\|)\}_{i\in\mathbb{N}}$ converge to $\alpha_{\infty\infty}$ Consequently $\{\chi(\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\}_{i\in\mathbb{N}}$ converges to 0 and, having fixed $\varepsilon > 0$, we can consider the first level n for which $\chi\left(\mathcal{A}_n^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right) \leq \varepsilon$. It then trivially verifies the inequality $\chi\left(\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right) > \varepsilon$, when $i < n$.

To prove now that $\chi\left(\mathcal{A}_i^\pi\left[\mathcal{D}_\sigma^\mathcal{K}\right]\right) \leq \varepsilon, \forall i \geq n$, it is enough to take into account that, by construction, the sequence $\{\mathcal{A}_i^{\pi'}\left[\mathcal{D}_\sigma^\mathcal{K}\right](\|\mathcal{D}_i\|)\}_{i\in\mathbb{N}}$ is monotonically decreasing and the trends are particular configurations of the same accuracy pattern. ■

Intuitively, this result determines the level from which the trends to be calculated estimate the final accuracy with a gain, on the current approximation, below a given threshold. As the layers converge strictly decreasing to 0 in synchrony with the uniform convergence of the learning trace, they introduce an alternative to absolute error and convergence thresholds, which in this case have proven to be impracticable. All these make it possible for us to fully exploit the model of convergence, giving practical sense to the proposal. Nonetheless, the real nature of the observations obliges us to keep in mind some additional considerations.

4.2. Robustness

Although the previous results certify that our proposal is theoretically correct with respect to its original specification, ensuring the termination in all cases, its practical application is subject to conditions that could affect this correctness. In particular, our abstract model has been built on an ideal conceptualization. This implies the assumption of a series of formal properties for the learning curves, although they may slightly diverge from this modelization. The problem moves then away from our working hypotheses, making it necessary to assess how far the methodology described is engaged. For that, we henceforth assume that any learning curve $\mathcal{A}_{\infty}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ is positive definite and upper bounded by 100, two conditions whose compliance we can always guarantee, but only quasi-strictly increasing and concave. These premises are now our *testing hypotheses*, which capture the notion of irregular observation.

The new conditions can alter the monotony of the asymptotic backbone, translating it into a quasi-monotony. This is not a minor problem because we are talking about the key to proving the uniform convergence in a learning trace, while this type of disorders generates only local disturbs and does not affect the convergence itself. The question then focuses on how to reduce the impact on the proximity criterion associated to the correctness, as is the case for the layers of convergence. For our study, we distinguish two types of alterations, according to their position in relation to the *working level*, namely the instance from which these possible dysfunctions would have an acceptable impact. Unfortunately, its optimal location depends on unpredictable factors such as the magnitude, evolution and the very existence of these disorders. Consequently, a theoretically well-founded characterization of this level is impossible and the formalization of heuristic criteria is the only way out.

Definition 8 *Let* $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ *be a learning trace with asymptotic backbone* $\{\alpha_i\}_{i\in\mathbb{N}}$, $\nu \in (0,1)$, $\varsigma \in \mathbb{N}$ *and* $\lambda \in \mathbb{N} \cup \{0\}$. *We define the working level for* $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ *with verticality threshold* ν, *slow down* ς *and look-ahead* λ, *as the smallest* $\omega(\nu,\varsigma,\lambda) \in \mathbb{N}$ *verifying*

$$\frac{\sqrt[\varsigma]{\nu}}{1-\nu} \geq \frac{|\alpha_{i+1}-\alpha_i|}{\|\mathcal{D}_{i+1}\|-\|\mathcal{D}_i\|}, \quad \forall i \in \mathbb{N} \quad \textit{such that} \quad \omega(\nu,\varsigma,\lambda) \leq i \leq \omega(\nu,\varsigma,\lambda)+\lambda \tag{14}$$

The smallest $\wp(\nu,\varsigma,\lambda) \geq \omega(\nu,\varsigma,\lambda)$ *with* $\alpha_{\wp(\nu,\varsigma,\lambda)} \leq 100$ *is the* prediction level for $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$.

We therefore determine the working level from the normalization $\nu \in (0,1)$ of the maximum permissible absolute value for the slope of the straight line joining two consecutive points on the asymptotic backbone. Since such values decrease as the training advances and the irregularities are inversely proportional to the degree of learning achieved, both magnitudes correlate. This allows us to reasonably categorize such alterations, using an optional extra degree of verification given by the look-ahead window. Retaking our example, the difference between the alterations in the monotony of the asymptotic backbone before and after the working level can be seen in the left-most diagram of Fig. 5, with $\nu = 2*10^{-5}$, $\varsigma = 1$ and $\lambda = 5$ as parameters. Furthermore, although the condition on the upper bound for the prediction level does not have any impact on the convergence we are studying, it allows us to focus on those learning

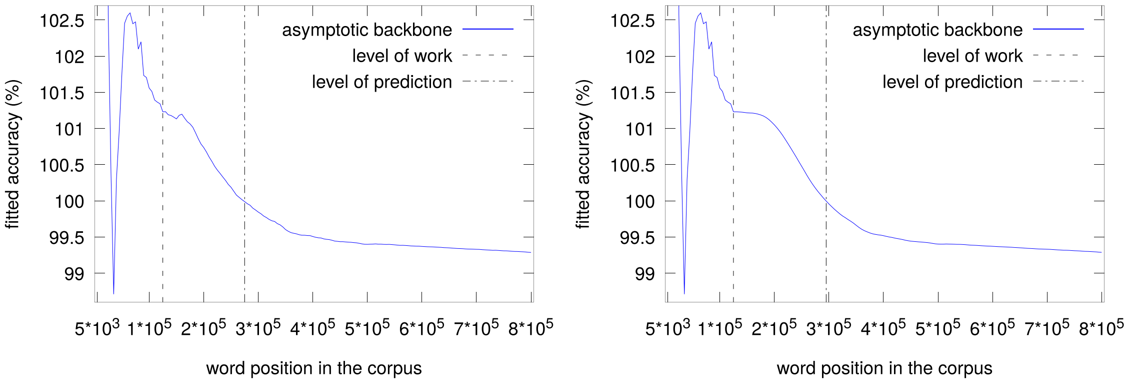

Fig. 5. Working and prediction levels for asymptotic backbones built without and with anchors for fnTBL on Frown.

trends that could be considered as feasible approximations for accuracy, whose highest value is 100. In practice, the working level is often aligned with the prediction one because the location of the former usually permits us to gain the time needed to stabilize the convergence process below such a maximum. Trivially, when the working hypotheses are fulfilled, the condition characterizing the working level verifies from the first operative level $\ell = 3$, whatever the parameters considered.

4.2.1. Irregularities before the working level

The low number of observations available, combined with the fact that they are associated with steep slopes, has here a multiplying effect on the irregularities on the asymptotic backbone, causing wide fluctuations. The use of large enough sets of observations to generate the learning trends should help to mitigate the problem, which is formally equivalent to estimating the working level. That is, the only effective strategy in this case to avoid alterations in the monotony of the asymptotic backbone is to discard all the trends associated to pre-working levels, as can be seen from the left-most diagram of Fig. 5.

4.2.2. Irregularities after the working level

The irregularities after the working level should be less pronounced, below the verticality threshold chosen, unless they are caused by a radical and continued shift away from the working hypotheses. In this case, the monotony of the asymptotic backbone can even be reversed until the training process allows the system to rebalance it. We are then talking about a problem outside our testing scenario, because its treatment requires the resampling of the training data. Otherwise, when we deal with small variations in the monotony, the access to an adequate extra observation could facilitate the realignment of the asymptotic backbone. We must here take into account that, the learning trends being the result of a fitting process, the impact of that additional example will be greater the further the associated instance is. In order to formalize this idea, we first extend the notion of learning trace.

Definition 9 *Let* $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ *be a learning trace with working level* ω(ν, ς, λ), *and* $\{\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](\infty)\}_{\ell>\omega(\nu,\varsigma,\lambda)} \subset \mathbb{R}^{+}$ *a sequence. A* learning trend of level $\ell > \omega(\nu, \varsigma, \lambda)$ with anchor $\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](\infty)$ for $\mathcal{A}_{\infty}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ using the accuracy pattern π, *is a curve* $\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] \in \pi$ *fitting the observations* $\{[x_i, \mathcal{A}_{\infty}[\mathcal{D}_{\sigma}^{\mathcal{K}}](x_i)], x_i := \|\mathcal{D}_i\|\}_{i=1}^{\ell} \cup [\infty, \hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](\infty)]$ *We then denote by* $\hat{\rho}_{\ell}(i) := [\mathcal{A}_{\infty}[\mathcal{D}_{\sigma}^{\mathcal{K}}] - \hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]](\|\mathcal{D}_i\|)$ *the* residual of $\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ at the level $i \in \mathbb{N}$, *by* $\hat{\rho}_{\ell}(\infty) := \hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](\infty) - \hat{\alpha}_{\ell}$ *its* residual at the point of infinity *and by* $y = \hat{\alpha}_{\ell}$ *its asymptote.*
When $\{\hat{\alpha}_{\ell}\}_{\ell>\omega(\nu,\varsigma,\lambda)}$ *is positive definite and converges monotonically to the asymptotic value* α_{∞} *of* $\mathcal{A}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$*, we say that* $\hat{\mathcal{A}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] := \{\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]\}_{\ell>\omega(\nu,\varsigma,\lambda)}$ *is an* anchoring learning trace, *whose* reference *is* $[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \omega(\nu, \varsigma, \lambda)]$ *with* $\{\hat{\alpha}_{\ell}\}_{\ell>\omega(\nu,\varsigma,\lambda)}$ *its* asymptotic backbone.

The only difference between an anchoring learning trace and a standard one is the use of extra fitting points in the infinity in order to generate its learning trends. In particular, as the properties of monotony and convergence for the asymptotic backbone must be preserved, the previous findings on uniform convergence, correctness and layer correctness for learning traces also verify. Accordingly, the conclusions and even the formal proofs for Theorems 1, 2 and 3 remain fully valid when we use anchors, which is why we do not repeat them. At this point, the next step in achieving a mechanism to soften the impact of irregular observations in the asymptotic backbone is to determine the real potential of an anchor in this regard.

Theorem 4 *Let* $\hat{\mathcal{A}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ *be an anchoring learning trace. We have* $\hat{\rho}_i(\infty) = -\sum_{j\leq i} \hat{\rho}_i(j), \forall i > \omega(\nu, \varsigma, \lambda)$ *and* $\lim_{i\to\infty} \hat{\rho}_i(\infty) = 0$.

Proof: Let us assume $\{\hat{\alpha}_i\}_{i\in\mathbb{N}}$ is the asymptotic backbone for $\hat{\mathcal{A}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$. Having fixed a level $i > \omega(\nu, \varsigma, \lambda)$, the fitting mechanism ensures that $\sum_{j\leq i} \hat{\rho}_i(j) + \hat{\rho}_i(\infty) = 0$ because the sum total of residuals on $\hat{\mathcal{A}}_{i}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ must be null, which proves the first part of the wording. As $\hat{\rho}_i(\infty) := \mathcal{A}_{i}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](\infty) - \hat{\alpha}_i, \forall i > \omega(\nu, \varsigma, \lambda)$, with both sequences $\{\hat{\mathcal{A}}_{i}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](\infty)\}_{i>\omega(\nu,\varsigma,\lambda)}$ and $\{\hat{\alpha}_i\}_{i>\omega(\nu,\varsigma,\lambda)}$ converging to α_{∞}, we conclude that $\lim_{i\to\infty} \hat{\rho}_i(\infty) = 0$. ■

This suggests that an intelligent use of anchors can reduce the magnitude of distortions in the asymptotic backbone, by compensating for their associated residuals. Thus, having fixed a learning trend, the degree of smoothing applicable to the irregularities correlates with its residual at the point of infinity. In other words, the better the estimation of the asymptotes by the anchors, the lower the impact of inconsistencies on the monotony of the asymptotic backbone. The question now is, therefore, how to choose the optimal sequence of anchors. Considering that the abstract model is iterative, our approach should take into account the impact of the new observation incorporated at each cycle, which excludes strategies based on static analysis.

Theorem 5 *Let $\hat{\mathcal{A}}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ be an anchoring learning trace with asymptotic backbone $\{\hat{\alpha}_i\}_{i>\omega(\nu,\varsigma,\lambda)}$. We then have:*

$$\forall i > \omega(\nu,\varsigma,\lambda),\ \hat{\mathcal{A}}^{\pi}_{i+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \leq \hat{\alpha}_i - \sum_{j\leq i+1} \hat{\rho}_{i+1}(j) \ \left(\text{resp. } \hat{\mathcal{A}}^{\pi}_{i+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \geq \hat{\alpha}_i - \sum_{j\leq i+1} \hat{\rho}_{i+1}(j)\right) \tag{15}$$

if $\{\hat{\alpha}_i\}_{i>\omega(\nu,\varsigma,\lambda)}$ is decreasing (resp. *increasing*).

Proof: In the decreasing case, we have $\hat{\alpha}_{i+1} \leq \hat{\alpha}_i, \forall i > \omega(\nu,\varsigma,\lambda)$. As $\hat{\mathcal{A}}^{\pi}_{i+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) - \hat{\alpha}_{i+1} := \hat{\rho}_{i+1}(\infty) = -\sum_{j\leq i+1} \hat{\rho}_{i+1}(j), \forall i > \omega(\nu,\varsigma,\lambda)$, the thesis derives immediately. The increasing case is analogous. ■

Since Theorem 4 stated that, excluding the one at the point of infinity, the residuals accumulated by the observations of the trends in an anchoring learning trace converge to zero, this last result opens the doors to a systematic way to generate anchors from the last asymptotic values calculated.

Theorem 6 *Let $\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ be a learning trace with asymptotic backbone $\{\alpha_i\}_{i\in\mathbb{N}}$, and $\left\{\hat{\mathcal{A}}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty)\right\}_{i>\omega(\nu,\varsigma,\lambda)}$ be the sequence defined by*

$$\hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) := \alpha_{\omega(\nu,\varsigma,\lambda)} \qquad \hat{\mathcal{A}}^{\pi}_{i+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) := \hat{\alpha}_i := \lim_{x\to\infty} \hat{\mathcal{A}}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x) \tag{16}$$

with $\hat{\mathcal{A}}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ a curve fitting $\{[x_j, \mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x_j)], x_j := \|\mathcal{D}_j\|\}^{i}_{j=1} \cup \left[\infty, \mathcal{A}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty)\right]$ $\forall i > \omega(\nu,\varsigma,\lambda)$. Then $\alpha_{\omega(\nu,\varsigma,\lambda)+i} \leq \hat{\alpha}_{\omega(\nu,\varsigma,\lambda)+i}$ (resp. $\alpha_{\omega(\nu,\varsigma,\lambda)+i} \geq \hat{\alpha}_{\omega(\nu,\varsigma,\lambda)+i}$), $\forall i \in \mathbb{N}$, *when $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing* (resp. *increasing*). *Furthermore, $\left\{\hat{\mathcal{A}}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}]\right\}_{i>\omega(\nu,\varsigma,\lambda)}$ is an anchoring learning trace of reference $\left[\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}], \omega(\nu,\varsigma,\lambda)\right]$. We call $\left\{\hat{\mathcal{A}}^{\pi}_{i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty)\right\}_{i>\omega(\nu,\varsigma,\lambda)}$ the set of* canonical anchors *for $\hat{\mathcal{A}}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$.*

Proof: As both the decreasing and the increasing cases are analogous, we only detail the former. We first demonstrate both $\alpha_{\omega(\nu,\varsigma,\lambda)+i} \leq \hat{\alpha}_{\omega(\nu,\varsigma,\lambda)+i}$ and $\hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+i+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \leq \hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty), \forall i \in \mathbb{N}$. We do so by induction on *i*. Let us first assume $i = 1$, as $\hat{\alpha}_{\omega(\nu,\varsigma,\lambda)+1} := \lim_{x\to\infty} \hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)$ and the latter is a curve fitting the set of values

$$\{[x_j, \mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x_j)], x_j := \|\mathcal{D}_j\|\}^{\omega(\nu,\varsigma,\lambda)+1}_{j=1} \cup \left\{\left[\infty, \mathcal{A}^{\pi}_{\omega(\nu,\varsigma,\lambda)+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty)\right]\right\} \tag{17}$$

or, in other words, the series

$$\{[x_j, \mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x_j)], x_j := \|\mathcal{D}_j\|\}^{\omega(\nu,\varsigma,\lambda)+1}_{j=1} \cup \{[\infty, \alpha_{\omega(\nu,\varsigma,\lambda)}]\} \tag{18}$$

where, as $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing, we have $\lim_{x\to\infty} \mathcal{A}^{\pi}_{\omega(\nu,\varsigma,\lambda)+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x) := \alpha_{\omega(\nu,\varsigma,\lambda)+1} \leq \alpha_{\omega(\nu,\varsigma,\lambda)}$ and we immediately deduce

$$\begin{aligned} \alpha_{\omega(\nu,\varsigma,\lambda)+1} &\leq \lim_{x\to\infty} \hat{\mathcal{A}}_{\omega(\nu,\varsigma,\lambda)+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x) := \hat{\alpha}_{\omega(\nu,\varsigma,\lambda)+1} := \hat{\mathcal{A}}_{\omega(\nu,\varsigma,\lambda)+2}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \leq \\ &\leq \alpha_{\omega(\nu,\varsigma,\lambda)} := \hat{\mathcal{A}}_{\omega(\nu,\varsigma,\lambda)+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \end{aligned} \tag{19}$$

which proves this first case.

Let us now assume that $\hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+i+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \leq \hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+i}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty), \forall i < n \in \mathbb{N}$. We must now prove that $\hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+n+1}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty) \leq \hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+n}[\mathcal{D}^{\mathcal{K}}_{\sigma}](\infty)$.

As $\hat{\alpha}_{\omega(\nu,\varsigma,\lambda)+n} := \lim_{x\to\infty} \hat{\mathcal{A}}^{\pi}_{\omega(\nu,\varsigma,\lambda)+n}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](x)$ and the latter is a curve fitting the set of values

$$\left\{\left[x_j, \mathcal{A}_{\infty}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](x_j)\right], x_j := \|\mathcal{D}_j\|\right\}_{j=1}^{\omega(\nu,\varsigma,\lambda)+n} \cup \left\{\left[\infty, \mathcal{A}^{\pi}_{\omega(\nu,\varsigma,\lambda)+n}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](\infty)\right]\right\} \tag{20}$$

or, in other words, the series

$$\left\{\left[x_j, \mathcal{A}_{\infty}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](x_j)\right], x_j := \|\mathcal{D}_j\|\right\}_{j=1}^{\omega(\nu,\varsigma,\lambda)+n} \cup \left\{\left[\infty, \alpha_{\omega(\nu,\varsigma,\lambda)+n-1}\right]\right\} \tag{21}$$

where, as $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing, we have $\lim_{x\to\infty} \mathcal{A}^{\pi}_{\omega(\nu,\varsigma,\lambda)+n}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](x) := \alpha_{\omega(\nu,\varsigma,\lambda)+n} \leq \alpha_{\omega(\nu,\varsigma,\lambda)+n-1}$. Since, by induction hypothesis, it also verifies that $\alpha_{\omega(\nu,\varsigma,\lambda)+n-1} \leq \alpha_{\omega(\nu,\varsigma,\lambda)+n-1}$, we immediately deduce

$$\begin{aligned} \alpha_{\omega(\nu,\varsigma,\lambda)+n-1} \quad &\leq \quad \lim_{x\to\infty} \mathcal{A}_{\omega(\nu,\varsigma,\lambda)+n}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](x) := \alpha_{\omega(\nu,\varsigma,\lambda)+n} := \mathcal{A}_{\omega(\nu,\varsigma,\lambda)+n+1}\left(\mathcal{D}^{\mathcal{K}}_{\sigma}\right)(\infty) \quad \leq \\ &\leq \quad \alpha_{\omega(\nu,\varsigma,\lambda)+n-1} := \mathcal{A}_{\omega(\nu,\varsigma,\lambda)+n}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](\infty) \end{aligned} \tag{22}$$

which proves that $\{\hat{\alpha}_i\}_{i>\omega(\nu,\varsigma,\lambda)}$ is, as the reference asymptotic backbone $\{\alpha_i\}_{i\in\mathbb{N}}$, positive definite and monotonically decreasing. Given that it is a infimum for this sequence, the latter converges to α_{∞} and the thesis is proven. ■

Regardless of the consideration of other types of anchors, we are already in a position to transfer their practical use for smoothing small irregularities, as should be the case once passed the working level. Obviously, faced with the impossibility of representing the value ∞ on the computer, we locate the anchors in a case as far as possible for practical purposes. Returning to our running example and identifying such a case with the word position 10^{200}, this technique is illustrated in Fig. 5, where the effects of using anchors in its right-most diagram are in contrast with the results obtained in their absence and shown in the left-most one. The result also shows that, as might have been expected, anchors can cause changes in the speed of convergence of the learning trace. In fact, given that these structures are necessarily estimated from the experience provided by previously computed learning trends, anchoring strategies tend to be conservative. In practice, this has meant the slowing down of the process, as has just been proven for the canonical case.

5. The testing frame

Given a training data base $\mathcal{D}$, the goal is to illustrate how far in advance and how well a learning curve $\mathcal{A}_{\infty}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right]$, built from a kernel $\mathcal{K}$ and using a step function σ, can be approximated. We apply our proposal based on the convergence of learning traces, without and with anchors.

5.1. The monitoring structure

As evaluation basis we introduce the *run*, a tuple $\mathcal{E} = \left[\mathcal{A}^{\pi}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right], \wp(\nu,\varsigma,\lambda), \tau\right]$ characterized by a learning trace $\mathcal{A}^{\pi}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right]$, a prediction level $\wp(\nu,\varsigma,\lambda)$ and a convergence threshold τ. We then formalize our experimental study on two collections of runs, $\mathcal{F} = \{\mathcal{E}_i\}_{i\in I}$ and $\mathcal{G} = \{\mathcal{E}_j\}_{j\in J}$, of standard and anchoring learning traces, respectively. The anchors are located in a case 10^{200}, sufficiently far for giving practical sense to the concept. With the aim of avoiding distortions due to the lack of uniformity in the testing frame, a common kernel $\mathcal{K}$, accuracy pattern π, step function σ, verticality threshold ν, slow down ς and look-ahead λ are used. In order to make the results comparable, we also use a common value for the convergence threshold τ on those runs involving the same training data base and system tested, whether they are in $\mathcal{F}$ or $\mathcal{G}$.

In practice, we are interested in studying each run $\mathcal{E} = \left[\mathcal{A}^{\pi}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right], \wp(\nu,\varsigma,\lambda), \tau\right]$ from the level in which predictions are below the convergence threshold τ and which we call *convergence level* (cLevel) from now on. So, once the prediction level (pLevel) is found during the computation of the learning trace $\mathcal{A}^{\pi}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right]$, we begin to check the layer of convergence. When it reaches the convergence threshold τ, the corresponding trend $\mathcal{A}^{\pi}_{\text{clevel}}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right]$ is selected for predicting the learning curve $\mathcal{A}_{\infty}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right]$, which implies that the iterative process of approximation is stopped at that instant. For the collection of runs $\mathcal{F} = \{\mathcal{E}_i\}_{i\in I}$ (resp. $\mathcal{G} = \{\mathcal{E}_j\}_{j\in J}$), monitoring is then applied to those learning trends

$\{\mathcal{A}^{\pi}_{\text{clevel}_i}[\mathcal{D}^{\mathcal{K}}_{\sigma}]\}_{i\in I}$ (resp. $\{\mathcal{A}^{\pi}_{\text{clevel}_j}[\mathcal{D}^{\mathcal{K}}_{\sigma}]\}_{j\in J}$) on a common set of *control levels* for the training data base, which are extracted from a finite sub-interval of the *prediction windows* $\{[\text{CLevel}_i, \infty)\}_{i\in I}$ (resp. $\{[\text{CLevel}_j, \infty)\}_{j\in J}$). We baptize these sets as *control sequences* and in each of their levels the accuracy (Ac) and its estimate (EAc) are computed for each run. In order to guarantee the soundness of the results, we use 6 decimal digits, though only two are represented for reasons of space and visibility.

5.2. *The quality metrics*

We want to measure both the reliability of our estimates for accuracy regarding the actual values and their robustness against variations in our working hypotheses, in all case studies. For that, two groups of metrics are used.

5.2.1. *Measuring the reliability*

A way of measuring the reliability is through the *mean absolute percent error* (MAPE) (Vandome, 1963). For every run $\mathcal{E}$ and level i of a control sequence $\mathcal{S}$, we first compute the *percentage error* (PE) as the difference between the EAc calculated from $\mathcal{A}^{\pi}_{\text{CLevel}_{\mathcal{E}}}[\mathcal{D}^{\mathcal{K}}_{\sigma}](i)$ and the Ac from $\mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](i)$. We can then express the MAPE as the arithmetic mean of the unsigned PE, as

$$\text{PE}(\mathcal{E})(i) := 100 * \frac{[\mathcal{A}^{\pi}_{\text{CLevel}_{\mathcal{E}}} - \mathcal{A}_{\infty}][\mathcal{D}^{\mathcal{K}}_{\sigma}](i)}{\mathcal{A}_{\infty}[\mathcal{D}^{\mathcal{K}}_{\sigma}](i)}, \mathcal{E} = [\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}], \wp(\nu, \varsigma, \lambda), \tau], i \in \mathcal{S} \tag{23}$$

$$\text{MAPE}(\mathcal{E})(\mathcal{S}) := \frac{100}{\|\mathcal{S}\|} * \sum_{i\in\mathcal{S}} |\text{PE}(\mathcal{E})(i)| \tag{24}$$

Intuitively, the error in the estimates done over a control sequence is, on average, proportional to the MAPE. Nonetheless, while at first sight this measure could appear to be sufficient for our purposes, it only provides quantitative information about the estimation process. We also need to determine to what extent those deviations impact decision-making on accuracy-based criteria, which is a qualitative perspective.

In this context, once a set of runs $\mathcal{H}$ working on a common training data base have been fixed, we are interested in calculating the percentage of those for which such errors do not cause wrong decisions to be made in looking for the run with the best performance. To this end, having fixed a control sequence $\mathcal{S}$, the reliability of one of these runs in relation to the others depends on its estimates not altering the relative position of its learning curve with respect to the learning curves associated to the rest of runs throughout $\mathcal{S}$. We then say that such run is *locally reliable in* $\mathcal{H}$ *along* $\mathcal{S}$. From this perspective, the fact that different runs have similar MAPE values may help to improve the reliability of one from the other, but only when the corresponding asymptotic backbones have the same type of monotony.[4] The reason is that the learning trends used for predicting are then close to geometric translations of the learning curves, on the basis of a common shifting vector. Otherwise, the proximity between MAPEs does not allow conclusions to be drawn on reliability, which is unfortunately our case.

This justifies the need for a complementary evaluation view, independent of the MAPE concept and based on the local reliability condition, though taking into account that calling for compliance with the latter may be sometimes unrealistic. The reason is that it applies to an entire control sequence, which may be an excessive degree of requirement when comparing runs whose learning curves intersect within their domain. More specifically, the error in the estimate of an intersection point should be lower than the distance between its neighboring control levels, which is highly unlikely if such distance is short. Therefore, we are here more interested in assessing the *rate of distortion* introduced when comparing runs, understood as the percentage of control levels in which the estimates of a run preserve the relative position of its learning curve with respect to the other ones associated with the rest of the runs.

Accordingly, we distinguish two testing scenarios from runs in $\mathcal{F}$ and $\mathcal{G}$. The former refers to the compliance for the local reliability condition when the corresponding learning curves are disjoint. The latter analyzes the impact of prediction errors in the comparison of runs whose learning curves intersect. Given a control sequence $\mathcal{S}$ and a set $\mathcal{H}$

[4] This implies that the learning traces approximate the learning curves from a common relative position to all the runs.

of runs defined on the same training data base, our primary reference in both scenarios is the *reliability estimation* (RE) *of two runs* $\mathcal{E}, \tilde{\mathcal{E}} \in \mathcal{H}$ *on* $i \in \mathcal{S}$, defined as

$$\mathrm{RE}\left(\mathcal{E}, \tilde{\mathcal{E}}\right)(i) := \begin{cases} 1 & \text{if } \left[\left[\mathcal{A}_{\infty} - \tilde{\mathcal{A}}_{\infty}\right] * \left[\mathcal{A}^{\pi}_{\mathrm{CLevel}_{\mathcal{E}}} - \tilde{\mathcal{A}}^{\pi}_{\mathrm{CLevel}_{\tilde{\mathcal{E}}}}\right]\right]\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right](i) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

with $\mathcal{E} = \left[\mathcal{A}^{\pi}\left[\mathcal{D}^{\mathcal{K}}_{\sigma}\right], \wp(\nu, \varsigma, \lambda), \tau\right]$, $\tilde{\mathcal{E}} = \left[\tilde{\mathcal{A}}^{\pi}\left[\mathcal{D}^{\tilde{\mathcal{K}}}_{\tilde{\sigma}}\right], \wp\left(\tilde{\nu}, \tilde{\varsigma}, \tilde{\lambda}\right), \tilde{\tau}\right]$ and $\mathcal{E} \neq \tilde{\mathcal{E}}$. Having fixed a control level, this Boolean function verifies if the estimates for $\mathcal{E}$ and $\tilde{\mathcal{E}}$ preserve the relative positions of the corresponding observations, and we can extend it to the control sequence.

Definition 10 *Let* $\mathcal{E}$ *and* $\tilde{\mathcal{E}}$ *be runs on a control sequence* $\mathcal{S}$. *We define the reliability estimation ratio* (RER) *of* $\mathcal{E}$ *and* $\tilde{\mathcal{E}}$ *for* $\mathcal{S}$ *as*

$$\mathrm{RER}\left(\mathcal{E}, \tilde{\mathcal{E}}\right)(\mathcal{S}) := 100 * \frac{\sum_{i \in \mathcal{S}} \mathrm{RE}\left(\mathcal{E}, \tilde{\mathcal{E}}\right)(i)}{\|\mathcal{S}\|} \tag{26}$$

Although the RER covers our requirements to measure the performance in the second scenario, this is not the case for the first one, where we need to calculate the number of runs in a set $\mathcal{H}$ with regard to which the estimates for a given one $\mathcal{E}$ are reliable on the whole of the control sequence $\mathcal{S}$ considered.

Definition 11 *Let* $\mathcal{H} = \{\mathcal{E}_k\}_{k \in K}$ *and* $\mathcal{E} \notin \mathcal{H}$ *be a set of runs and a run, respectively, on a control sequence* $\mathcal{S}$. *We define the* decision-making reliability (DMR) of $\mathcal{E}$ on $\mathcal{H}$ for $\mathcal{S}$ *as*

$$\mathrm{DMR}(\mathcal{E}, \mathcal{H})(\mathcal{S}) := 100 * \frac{\|\mathcal{E}_k \in \mathcal{H}, \mathrm{RER}(\mathcal{E}, \mathcal{E}_k)(\mathcal{S}) = 100\|}{\|\mathcal{S}\|} \tag{27}$$

Together these metrics offer a good overview of the prediction reliability achieved. Thus, MAPE gives us a way to quantitatively evaluate our estimates regardless of the scenario considered, while DMR and RER provide a qualitative point of view. Once a set of runs defined on the same training data base has been fixed, DMR focuses on the reliability of one of these with respect to the rest along complete control sequences, while RER provides a more realistic view when comparing two runs involving intersecting learning curves.

5.2.2. Measuring the robustness

Since the stability of a run correlates to the degree of monotony in the asymptotic backbone, a way to measure it is to calculate the percentage of monotonic elements in the latter. We are only interested in those elements computed between the working and the convergence levels, because it is in this interval where the approximation performs effectively.

Definition 12 *Let* $\mathcal{E}$ *be a run with asymptotic backbone* $\{\alpha_\ell\}_{\ell \in \mathbb{N}}$, *and* $\mathrm{CLevel}_{\mathcal{E}}$ *and* $\mathrm{WLevel}_{\mathcal{E}}$ *its convergence and working levels, respectively. We define the robustness rate* (RR) of $\mathcal{E}$ *as*

$$\mathrm{RR}(\mathcal{E}) := 100 * \frac{\|\mu\|}{\|\{\alpha_i, \mathrm{WLevel}_{\mathcal{E}} \leq i \leq \mathrm{CLevel}_{\mathcal{E}}\}\|} \tag{28}$$

with μ *the longest maximum monotonic subsequence of* $\{\alpha_i, \mathrm{WLevel}_{\mathcal{E}} \leq i \leq \mathrm{CLevel}_{\mathcal{E}}\}$.

Having fixed a run, its tolerance for alterations in the working hypotheses will be greater, the higher its RR. This provides an efficient criterion for checking the degree of robustness on which we can count.

6. The experiments

We focus on the prediction of learning curves associated to ML-based tagger generation, a demanding task in the domain of NLP, to illustrate our proposal. To that end, we first introduce the linguistic resources and settings used to later analyze the results on the basis of the testing frame described.

6.1. The linguistics resources and settings

We need a collection of corpora as training data bases on a number of target languages, as well as tag-sets, taggers and a methodology to compute the accuracy values we use for both computing and evaluating predictions. As target languages we consider English and Spanish. The former is the most widely studied and best understood one. Spanish is one of the languages with the highest growth and development potential in NLP and, in contrast to English, it has a complex derivational paradigm.

6.1.1. The corpora

With regard to the corpora, we have selected them together with their associated tag-sets from the most popular ones in the domain for each target language:

(1) The AnCora (Taulé et al., 2008) treebank includes a section for Spanish, previously used as a resource in the shared tasks of CONNL (Hajič et al., 2009) and semEval (Recasens et al., 2010). It has served as a training and testing resource for POS tagging (Hulden and Francom, 2012), parsing (Popel et al., 2013) and semantic annotation (Mukund et al., 2010) tasks. Since its tag-set has been developed for languages morphologically richer than English, AnCora has the most detailed annotation of the corpora considered and is the only one to follow the Eagles recommendations (Monachini and Calzolari, 1996). Its 280 tags (Taulé et al., 2008) cover the main POS classes used in Spanish as well as sub-classes and morphological features, accessible on clic.ub.edu/corpus/webfm_send/18. The number of words in the Spanish section of the treebank is more than 515,000.

(2) The Freiburg-Brown of American English (Mair and Leech, 2007) (FROWN) matches the composition and style of the Brown corpus (Francis and Kučera, 1967). It has been used in linguistic studies with a more theoretical purpose (Leech, 2009; Mair, 2006), which in our opinion entails an interesting counterpoint to the other two corpora considered, which are more oriented to NLP-related applications. The associated tag-set is the UCREL C8. With 169 tags accessible on ucrel.lancs.ac.uk/claws8tags.pdf, it was selected as the common tag-set for the Brown family of corpora (Hinrichs et al., 2010). The size of the Frown corpus is more than 1,165,000 words.

(3) The section with news items from the *Wall Street Journal* (WSJ) included in the Penn treebank (Marcus et al., 1999) is a popular corpus in NLP for both POS tagging (Brants, 2000; Collins, 2002; Giménez and Márquez, 2004; Ratnaparkhi, 1996; Toutanova et al., 2003) and parsing purposes (Charniak, 2000; Petrov et al., 2006). It has also been used in the shared tasks of prestigious events in the domain of natural language learning, such as CONNL (Hajič et al., 2009; Nivre et al., 2007; Surdeanu et al., 2008), and semantic evaluation, such as semEval (Yuret et al., 2010). The tag-set associated to this corpus has 45 tags, accessible on www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html and covering basic POS classes in English along with some morphological information. This is the simplest of the tag-sets we considered. The WSJ section is the biggest of our corpora with more than 1,170,000 words.

Both Penn and AnCora are treebanks annotated with POS tags as well as syntactic structures. By stripping them of the latter, they can be used to train POS tagging systems. In order to ensure well-balanced corpora, we also have scrambled them at sentence level before testing. Thus, the levels in the learning traces refer to word positions in the scrambled versions of AnCora, Frown and Penn.

6.1.2. The POS tagging systems

We avoid rule-based taggers, where the absence of a training phase leaves the consideration of accuracy as a predictable magnitude void of content. We then focus on systems built from supervised learning. In contrast to taggers resulting from unsupervised techniques, these make it possible to work with predefined tag-sets, which facilitates both the evaluation and the comprehension of the results. Furthermore, they have proven to be the best-performing tagging proposals, placing them as a reference for any testing purpose. We have selected a broad range of proposals covering the most representative architectures:

(1) In the category of stochastic methods and as representative of the *hidden Márkov models* (HMMS), we have chosen TNT (Brants, 2000). We also include here the TREETAGGER (Schmid, 1994), which uses decision trees to generate the HMM, and MORFETTE (Chrupala et al., 2008), an averaged perceptron approach (Collins, 2002). To illustrate the *maximum entropy models* (MEMS), we chose to work with MXPOST (Ratnaparkhi, 1996) and the tagger associated to APACHE OPENNLP (OPENNLP MAXENT) (see opennlp.apache.org/). Finally, the stanford POS tagger (Toutanova et al., 2003) is based on a *conditional Márkov model*, which combines features of HMMS and MEMS.

(2) Under the heading of other POS tagging methods, the possibilities are many and various. As an example of transformation-based learning, we take FNTBL (Ngai and Florian, 2001), an updated version of the classic BRILL (Brill, 1995). In relation to memory-based learning, the representative is the *memory-based tagger* (MBT) (Daelemans et al., 1996), while we chose SVMTOOL (Giménez and Márquez, 2004) to describe the behavior with respect to a support vector machine technique. Finally, we use a perceptron-based training method with look-ahead, through LAPOS (Tsuruoka et al., 2011).

6.1.3. The testing space

In order to avoid learning dysfunctions resulting from sentence truncation, we take a particular class of learning scheme, allowing us to reap the maximum from the training process. So, given a corpus $\mathcal{D}$, a kernel $\mathcal{K} \subsetneq \mathcal{D}$ and a step function σ, we build the set of individuals $\{\mathcal{D}_i\}_{i \in \mathbb{N}}$ as follows:

$$\mathcal{D}_i := [\![\mathcal{C}_i]\!], \forall i \in \mathbb{N}, \text{such that } \mathcal{C}_1 := \mathcal{K} \text{ and } \mathcal{C}_i := \mathcal{C}_{i-1} \cup \mathcal{I}_i, \mathcal{I}_i \subset \mathcal{C} \setminus \mathcal{C}_{i-1}, \|\mathcal{I}_i\| := \sigma(i), \forall i \geq 2 \tag{29}$$

where $[\![\mathcal{C}_i]\!]$ denotes the minimal set of sentences including $\mathcal{C}_i$.

In this way, with respect to the setting of runs, the size of the kernels is $5 * 10^3$ words and the constant step function $5 * 10^3$ locates the instances. We deem both choices conservative, since much smaller kernels and increases are possible. In addition, we have found that reasonably good values for the prediction levels can be obtained with $\nu = 2 * 10^{-5}$, $\varsigma = 1$ and $\lambda = 5$. Following previous works on the performance of different models for fitting learning curves (Gu et al., 2001), particularly in the NLP domain (Kolachina et al., 2012), the power law family has been chosen for π. As regresion technique for approximating the partial learning curves we consider the *trust region method* (Branch et al., 1999).

Since we are trying to assess the validity or our proposal on finite corpora, it makes sense to study the prediction capacity within their boundaries. We then limit the scope in measuring the layer of convergence introduced in Definition 7 for a learning trend $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$. Formally, the asymptotic value α_i is replaced by the one reached at the end of the corpus we are working on. So, if $\lceil\!\lceil \ell \rceil\!\rceil$ denotes the position of the first sentence-ending beyond the ℓ-th word, the layer of convergence for $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ is now expressed as follows:

$$\mathcal{X}^\ell(\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]) := |\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}](\|\mathcal{D}_i\|) - \mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}](\lceil\!\lceil \ell \rceil\!\rceil)|$$

with $\ell = 5 * 10^5$ for runs involving the corpus ANCORA, and $\ell = 8 * 10^5$ for the rest. The term represented in bold font is the updated one.

The sampling window comprises the interval $[\lceil\!\lceil 5 * 10^3 \rceil\!\rceil, \lceil\!\lceil 5 * 10^5 \rceil\!\rceil]$ (resp. $[\lceil\!\lceil 5 * 10^3 \rceil\!\rceil, \lceil\!\lceil 8 * 10^5 \rceil\!\rceil]$) for ANCORA (resp. for FROWN and PENN), while the control levels are taken from control sequences in $[\lceil\!\lceil 3 * 10^5 \rceil\!\rceil, \lceil\!\lceil 5 * 10^5 \rceil\!\rceil]$ (resp. $[\lceil\!\lceil 3 * 10^5 \rceil\!\rceil, \lceil\!\lceil 8 * 10^5 \rceil\!\rceil]$), matching the location for instances. With the goal of conferring additional stability on our measures, we opt for a k-fold cross validation (Clark et al., 2010) for computing the samples, due to its good adaptation to small data sets, an advantage in our context. We have chosen $k = 10$, which has been used before in POS tagging evaluation (Daelemans et al., 1996; Giesbrecht and Evert, 2009).

6.2. The analysis of the results

As mentioned, our experiments apply to two complementary points of view, quantitative and qualitative, according, in this latter case, to the presence or absence of intersection points between the learning curves involved in the analysis and generated from a common training data base.

6.2.1. The sets of runs

We start from two collections of runs, $\mathcal{F} = \{\mathcal{E}_i\}_{i \in I}$ and $\mathcal{G} = \{\mathcal{E}_j\}_{j \in J}$, in order to illustrate the predictability of standard and anchoring learning traces, respectively. Since the use of anchors has been designed as a mechanism to reinforce the robustness of the approximations, and Theorem 6 states that it can introduce a delay on the convergence, we are particularly interested in comparing the performance in both cases. We then consider runs in $\mathcal{G}$ that are *homologous* to those in $\mathcal{F}$, namely they are exclusively distinguishable by the use of anchors. The detail of the monitoring is compiled separately for each case study. So, the collection $\mathcal{F}$ (resp. $\mathcal{G}$) of runs without (resp. with) anchors is shown in Table 1 (resp. Table 2), also including in passing the data relative to tests involving disjoint learning curves along the control sequences. Similarly, the data for the experiments involving intersecting learning curves for runs in $\mathcal{F}$ (resp. $\mathcal{G}$) can be seen in Table 3 (resp. Table 4). We include for each run the PLevel and the CLevel, as well as the values for Ac and EAc on some of the control levels considered to later calculate the results for MAPE, DMR and RR. In order to facilitate understanding, all those levels are indicated by their associated word positions in the corpus, which is denoted by a superscript **wp** in the identification labels.

These runs combine all the taggers and corpora previously introduced, though we have discarded some cases due to their high prediction level, which prevents us from evaluating them from the observations available. So, the run for TreeTagger on AnCora without (resp. with) anchors does not even reach this level on the interval covered by the sampling set, which is why it has not been included in the collection $\mathcal{F}$ (resp. $\mathcal{G}$) and its entry on Table 1 (resp. Table 2) is empty. In the case of the run for TreeTagger on Frown without (resp. with) anchors, the prediction level is 795,029 (resp. 800,010). This places us at the limit of our observations, therefore making an appropriate evaluation impossible, which also justifies its exclusion from collection $\mathcal{F}$ (resp. $\mathcal{G}$), as we can see in Table 1 (resp. Table 2).

Table 1
Monitoring of runs, without anchors, involving disjoint learning curves along the control sequences.

		PLEVEL^{wp}	τ	CLEVEL^{wp}	Control-Level^{wp}				MAPE	DMR	Control-Level^{wp}				MAPE	DMR	RR
					$[\![3*10^5]\!]$		$[\![5*10^5]\!]$				$[\![6*10^5]\!]$		$[\![8*10^5]\!]$				
					Ac	EAc	Ac	EAc			Ac	EAc	Ac	EAc			
fnTBL	AnCora	100,034	0.83	230,019	96.67	96.59	97.15	97.09	0.05	88.89							88.89
	Frown	275,023	1.50	290,002	94.98	95.09	95.78	95.90	0.14	100.00	95.98	96.15	96.31	96.52	0.16	100.00	88.24
	Penn	95,007	0.58	285,005	96.10	96.04	96.43	96.35	0.05	100.00	96.53	96.45	96.72	96.58	0.09	87.50	74.36
LAPOS	AnCora	60,019	0.36	280,005	97.55	97.47	97.90	97.78	0.10	100.00							88.89
	Frown	90,000	1.27	230,005	96.31	96.37	96.82	96.90	0.07	100.00	96.98	97.07	97.21	97.30	0.08	100.00	86.21
	Penn	65,003	0.93	140,002	96.81	96.83	97.07	97.06	0.01	100.00	97.15	97.13	97.28	97.23	0.02	100.00	50.00
MaxEnt	AnCora	70,009	1.07	155,053	96.23	96.02	96.77	96.41	0.30	77.78							72.22
	Frown	160,004	1.70	245,011	94.32	94.40	95.11	95.16	0.08	100.00	95.33	95.39	95.69	95.73	0.06	100.00	100.00
	Penn	95,007	0.60	270,033	95.95	95.88	96.34	96.18	0.11	100.00	96.45	96.27	96.63	96.40	0.17	100.00	100.00
MBT	AnCora	70,009	0.47	280,005	96.10	96.00	96.63	96.40	0.18	100.00							100.00
	Frown	215,030	1.95	255,003	93.58	93.61	94.52	94.51	0.04	100.00	94.77	94.80	95.17	95.22	0.04	100.00	100.00
	Penn	75,035	1.66	195,007	95.24	95.32	95.76	95.91	0.12	100.00	95.89	96.10	96.13	96.37	0.17	100.00	80.00
Morfette	AnCora	65,035	0.73	175,002	97.18	97.03	97.52	97.33	0.18	77.78							86.96
	Frown	100,009	1.43	240,053	95.65	95.65	96.23	96.27	0.04	100.00	96.39	96.47	96.69	96.75	0.06	100.00	51.72
	Penn	75,035	0.52	210,013	96.47	96.44	96.74	96.64	0.07	100.00	96.81	96.70	96.96	96.77	0.11	87.50	96.43
MXPOST	AnCora	80,023	1.15	205,013	96.56	96.57	97.08	97.15	0.04	88.89							42.31
	Frown	110,017	2.84	150,014	94.75	94.65	95.49	95.45	0.07	88.89	95.74	95.70	96.09	96.05	0.05	100.00	88.89
	Penn	85,013	1.40	140,002	96.11	96.18	96.52	96.54	0.04	100.00	96.59	96.64	96.74	96.79	0.04	100.00	100.00
Stanford	AnCora	40,010	0.51	255,008	96.86	96.72	97.31	97.09	0.18	88.89							93.18
	Frown	120,003	1.91	180,021	95.46	95.55	96.08	96.19	0.13	100.00	96.27	96.39	96.56	96.68	0.12	77.78	89.47
	Penn	90,031	0.98	150,031	96.41	96.35	96.72	96.62	0.08	100.00	96.81	96.70	96.95	96.82	0.11	100.00	53.85
SVMTool	AnCora	70,009	0.76	200,051	97.03	96.91	97.47	97.30	0.14	87.50							92.59
	Frown	205,004	1.41	260,002	95.77	95.82	96.37	96.50	0.12	100.00	96.54	96.70	96.79	97.01	0.15	100.00	100.00
	Penn	130,008	1.25	155,010	96.30	96.41	96.56	96.77	0.16	77.78	96.69	96.88	96.81	97.05	0.20	77.78	100.00
TnT	AnCora	60,019	0.77	175,002	97.11	97.03	97.47	97.35	0.09	100.00							62.50
	Frown	95,018	1.51	205,004	95.67	95.74	96.25	96.31	0.09	87.50	96.42	96.48	96.63	96.74	0.08	87.50	100.00
	Penn	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04	85.71	96.45	96.39	96.57	96.47	0.06	100.00	45.71
TreeTagger	AnCora		0.55		96.09		96.67										
	Frown	795,029	2.00		94.65		95.47				95.75		96.06				
	Penn	60,015	1.32	220,032	95.13	95.07	95.83	95.60	0.20	100.00	95.94	95.77	96.11	96.01	0.18	100.00	39.39

Table 2
Monitoring of runs, with anchors, involving disjoint learning curves along the control sequences.

		PLEVEL^{wp}	τ	CLEVEL^{wp}	Control-Level^{wp}				MAPE	DMR	Control-Level^{wp}				MAPE	DMR	RR
					$[\![3*10^5]\!]$		$[\![5*10^5]\!]$				$[\![6*10^5]\!]$		$[\![8*10^5]\!]$				
					Ac	EAc	Ac	EAc			Ac	EAc	Ac	EAc			
fnTBL	AnCora	100,034	0.83	230,019	96.67	96.58	97.15	97.08	0.06	88.89							100.00
	Frown	295,019	1.50	295,019	94.98	95.10	95.78	95.91	0.15	100.00	95.98	96.17	96.31	96.55	0.18	100.00	97.14
	Penn	95,007	0.58	285,005	96.10	96.04	96.43	96.35	0.05	100.00	96.53	96.44	96.72	96.58	0.09	87.50	100.00
LAPOS	AnCora	60,019	0.36	280,005	97.55	97.47	97.90	97.78	0.10	100.00							82.22
	Frown	90,000	1.27	230,005	96.31	96.37	96.82	96.91	0.08	100.00	96.98	97.07	97.21	97.31	0.09	100.00	72.41
	Penn	65,003	0.93	140,002	96.81	96.83	97.07	97.06	0.01	100.00	97.15	97.13	97.28	97.23	0.02	100.00	93.75
MaxEnt	AnCora	70,009	1.07	150,022	96.23	95.99	96.77	96.36	0.35	77.78							82.35
	Frown	160,004	1.70	250,001	94.32	94.41	95.11	95.17	0.09	100.00	95.33	95.41	95.69	95.76	0.08	100.00	100.00
	Penn	95,007	0.60	265,005	95.95	95.87	96.34	96.17	0.12	100.00	96.45	96.26	96.63	96.38	0.18	100.00	100.00
MBT	AnCora	70,009	0.47	280,005	96.10	95.99	96.63	96.39	0.19	100.00							100.00
	Frown	255,003	1.95	260,002	93.58	93.64	94.52	94.56	0.08	100.00	94.77	94.85	95.17	95.29	0.09	100.00	100.00
	Penn	75,035	1.66	175,019	95.24	95.18	95.76	95.72	0.06	100.00	95.89	95.89	96.13	96.13	0.03	100.00	90.48
Morfette	AnCora	65,035	0.73	175,002	97.18	97.02	97.52	97.32	0.19	77.78							95.65
	Frown	100,009	1.43	240,053	95.65	95.65	96.23	96.28	0.05	100.00	96.39	96.48	96.69	96.76	0.07	100.00	62.07
	Penn	75,035	0.52	210,013	96.47	96.44	96.74	96.64	0.07	100.00	96.81	96.70	96.96	96.77	0.12	87.50	100.00
MXPOST	AnCora	80,023	1.15	205,013	96.56	96.57	97.08	97.15	0.04	88.89							65.38
	Frown	110,017	2.84	150,014	94.75	94.66	95.49	95.47	0.05	88.89	95.74	95.71	96.09	96.06	0.04	100.00	100.00
	Penn	85,013	1.40	140,002	96.11	96.21	96.52	96.58	0.08	87.50	96.59	96.68	96.74	96.84	0.08	87.50	100.00
Stanford	AnCora	40,010	0.51	250,008	96.86	96.70	97.31	97.06	0.20	77.78							88.37
	Frown	145,014	1.91	190,034	95.46	95.59	96.08	96.26	0.19	77.78	96.27	96.46	96.56	96.77	0.19	77.78	95.24
	Penn	90,031	0.98	150,031	96.41	96.36	96.72	96.63	0.07	100.00	96.81	96.72	96.95	96.84	0.09	100.00	69.23
SVMTool	AnCora	70,009	0.76	200,051	97.03	96.90	97.47	97.28	0.16	87.50							100.00
	Frown	230,005	1.41	265,000	95.77	95.83	96.37	96.51	0.13	100.00	96.54	96.72	96.79	97.02	0.17	100.00	100.00
	Penn	130,008	1.25	160,001	96.30	96.45	96.56	96.83	0.20	77.78	96.69	96.95	96.81	97.12	0.26	77.78	100.00
TnT	AnCora	60,019	0.77	170,007	97.11	97.00	97.47	97.31	0.13	87.50							52.17
	Frown	95,018	1.51	210,009	95.67	95.77	96.25	96.35	0.13	87.50	96.42	96.53	96.63	96.79	0.12	87.50	100.00
	Penn	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04	85.71	96.45	96.38	96.57	96.47	0.07	100.00	48.57
TreeTagger	AnCora		0.55		96.09		96.67										
	Frown	800,010	2.00		94.65		95.47				95.75		96.06				
	Penn	60,015	1.32	220,032	95.13	95.06	95.83	95.59	0.21	100.00	95.94	95.76	96.11	95.99	0.19	100.00	69.70

Table 3
Monitoring of pairs of runs, without anchors, involving crossing learning curves along the control sequences.

		PLEVEL^{wp}	τ	CLEVEL^{wp}	Control-Level^{wp}				MAPE	DMR	Control-Level^{wp}				MAPE	DMR	RR
					$[\![3*10^5]\!]$		$[\![5*10^5]\!]$				$[\![6*10^5]\!]$		$[\![8*10^5]\!]$				
					Ac	EAc	Ac	EAc			Ac	EAc	Ac	EAc			
AnCora	SVMTool	70,009	0.76	200,051	97.03	96.91	97.47	97.30	0.14	92.68							92.59
	TnT	60,019	0.77	175,002	97.11	97.03	97.47	97.35	0.09								62.50
Frown	Morfette	100,009	1.43	240,053	95.65	95.65	96.23	96.27	0.04	80.49	96.39	96.47	96.69	96.75	0.06	79.21	51.72
	TnT	95,018	1.51	205,004	95.67	95.74	96.25	96.31	0.09		96.42	96.48	96.63	96.74	0.08		100.00
Penn	fnTBL	95,007	0.58	285,005	96.10	96.04	96.43	96.35	0.05	87.80	96.53	96.45	96.72	96.58	0.09	95.05	74.36
	TnT	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04		96.45	96.39	96.57	96.47	0.06		45.71
	MaxEnt	95,007	0.60	270,033	95.95	95.88	96.34	96.18	0.11		96.45	96.27	96.63	96.40	0.17	58.42	100.00
	TnT	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04		96.45	96.39	96.57	96.47	0.06		45.71
	MBT	75,035	1.66	195,007	95.24	95.32	95.76	95.91	0.12	31.71	95.89	96.10	96.13	96.37	0.17	37.62	80.00
	TreeTagger	60,015	1.32	220,032	95.13	95.07	95.83	95.60	0.20		95.94	95.77	96.11	96.01	0.18		39.39
	Morfette	75,035	0.52	210,013	96.47	96.44	96.74	96.64	0.07	97.56	96.81	96.70	96.96	96.77	0.11	77.23	96.43
	Stanford	90,031	0.98	150,031	96.41	96.35	96.72	96.62	0.08		96.81	96.70	96.95	96.82	0.11		53.85
	MXPOST	85,013	1.40	140,002	96.11	96.18	96.52	96.54	0.04	97.56	96.59	96.64	96.74	96.79	0.04	99.01	100.00
	TnT	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04		96.45	96.39	96.57	96.47	0.06		45.71

Table 4
Monitoring of pairs of runs, with anchors, involving crossing learning curves along the control sequences.

	PLEVEL^{wp}	τ	CLEVEL^{wp}	Control-Level^{wp}					MAPE	DMR	Control-Level^{wp}				MAPE	DMR	RR
				$[\![3*10^5]\!]$			$[\![5*10^5]\!]$				$[\![6*10^5]\!]$		$[\![8*10^5]\!]$				
					Ac	EAc	Ac	EAc			Ac	EAc	Ac	EAc			
AnCora	SVMTool	70,009	0.76	200,051	97.03	96.90	97.47	97.28	0.16	92.68							100.00
	TnT	60,019	0.77	170,007	97.11	97.00	97.47	97.31	0.13								52.17
Frown	Morfette	100,009	1.43	240,053	95.65	95.65	96.23	96.28	0.05	80.49	96.39	96.48	96.69	96.76	0.07	74.26	62.07
	TnT	95,018	1.51	210,009	95.67	95.77	96.25	96.35	0.13		96.42	96.53	96.63	96.79	0.12		100.00
Penn	fnTBL	95,007	0.58	285,005	96.10	96.04	96.43	96.35	0.05	87.80	96.53	96.44	96.72	96.58	0.09	95.05	100.00
	TnT	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04		96.45	96.38	96.57	96.47	0.07		48.57
	MaxEnt	95,007	0.60	265,005	95.95	95.87	96.34	96.17	0.12		96.45	96.26	96.63	96.38	0.18	58.42	100.00
	TnT	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04		96.45	96.38	96.57	96.47	0.07		48.57
	MBT	75,035	1.66	175,019	95.24	95.18	95.76	95.72	0.06	31.71	95.89	95.89	96.13	96.13	0.03	37.62	90.48
	TreeTagger	60,015	1.32	220,032	95.13	95.06	95.83	95.59	0.21		95.94	95.76	96.11	95.99	0.19		69.70
	Morfette	75,035	0.52	210,013	96.47	96.44	96.74	96.64	0.07	97.56	96.81	96.70	96.96	96.77	0.12	73.27	100.00
	stanford	90,031	0.98	150,031	96.41	96.36	96.72	96.63	0.07		96.81	96.72	96.95	96.84	0.09		69.23
	MXPOST	85,013	1.40	140,002	96.11	96.21	96.52	96.58	0.08	97.56	96.59	96.68	96.74	96.84	0.08	99.01	100.00
	TnT	60,015	0.51	230,002	96.13	96.10	96.39	96.32	0.04		96.45	96.38	96.57	96.47	0.07		48.57

Consequently, the entries in Tables 3 and 4 for the pairs in which those runs are involved are discarded. This is the case of MBT and TreeTagger on AnCora, and MXPOST and TreeTagger on Frown.

6.2.2. *The quantitative study*

Our reference performance metric is now the MAPE, and we are interested in values that are as small as possible. Regarding the collection of runs $\mathcal{F}$, when anchors are discarded, the results are shown in Fig. 6, from the data compiled in Table 1. Here, MAPES range from 0.01 (resp 0.02) for LAPOS on Penn, to 0.30 for MaxEnt on AnCora (resp. 0.20 for SVMTool on Penn) in the interval $\left[[\![3*10^5]\!],[\![5*10^5]\!]\right]$ (resp. $\left[[\![3*10^5]\!],[\![8*10^5]\!]\right]$). Those results are illustrated in Fig. 7, showing the learning curves and learning trends used for prediction on the runs with best and worst MAPE on both control sequences. As we have already done, the observations are generated considering the portion of the corpus taken from its beginning until the word position indicated on the horizontal axis. Finally, 57.14% (resp. 52.63%) of MAPE values in this set of runs belong to the interval [0,0.10], a proportion that reaches 96.43% (resp. 100%) in [0,0.20].

Focusing now on the collection of runs $\mathcal{G}$, when anchors are used, the results are shown in Fig. 8 from the data compiled in Table 2. This time the MAPES vary from 0.01 (resp. 0.02) for LAPOS on Penn, to 0.35 for MaxEnt on AnCora (resp. 0.26 for SVMTool on Penn) in the interval $\left[[\![3*10^5]\!],[\![5*10^5]\!]\right]$ (resp. $\left[[\![3*10^5]\!],[\![8*10^5]\!]\right]$). The trends involved are shown in Fig. 9, once again considering word positions in the text to indicate the portion from the beginning of

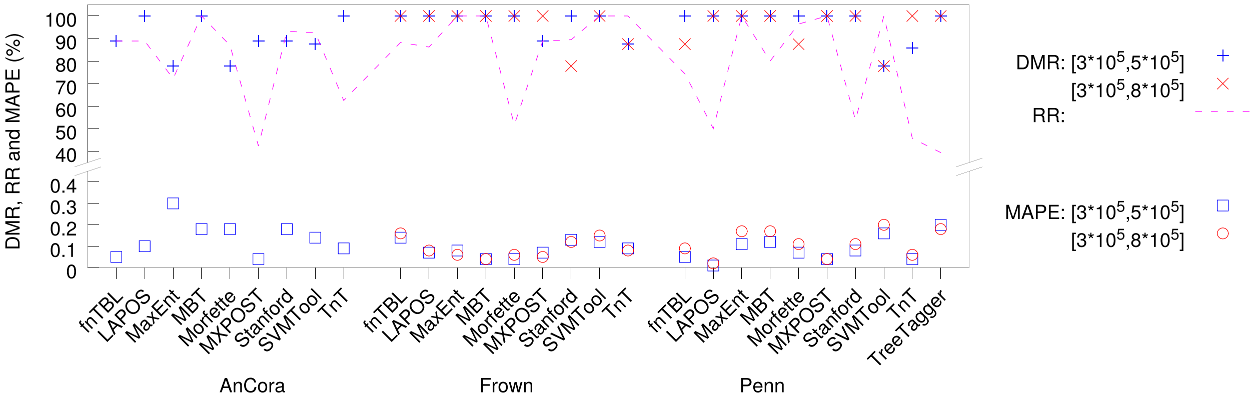

Fig. 6. MAPES and RRS for runs without anchors. DMRS when excluding crossing learning curves along the control sequences.

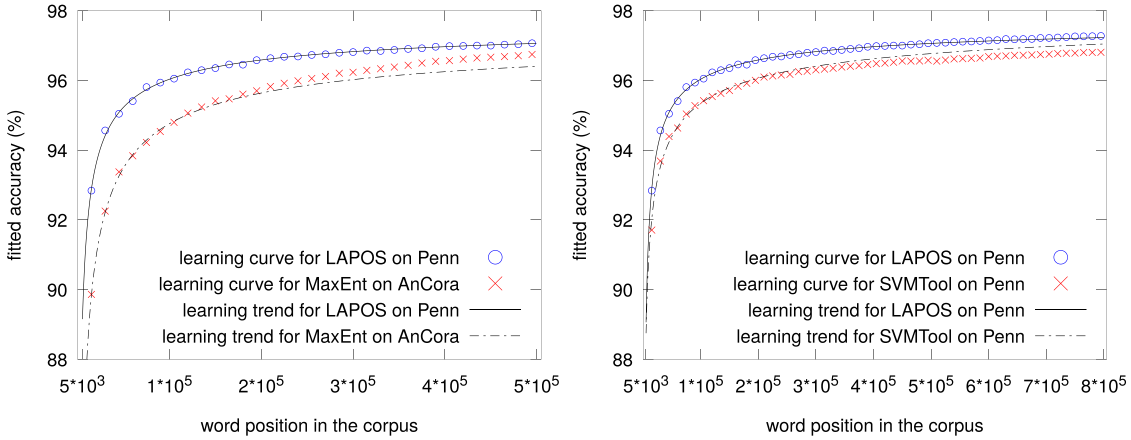

Fig. 7. Learning trends, without anchors, for the best and worst MAPES.

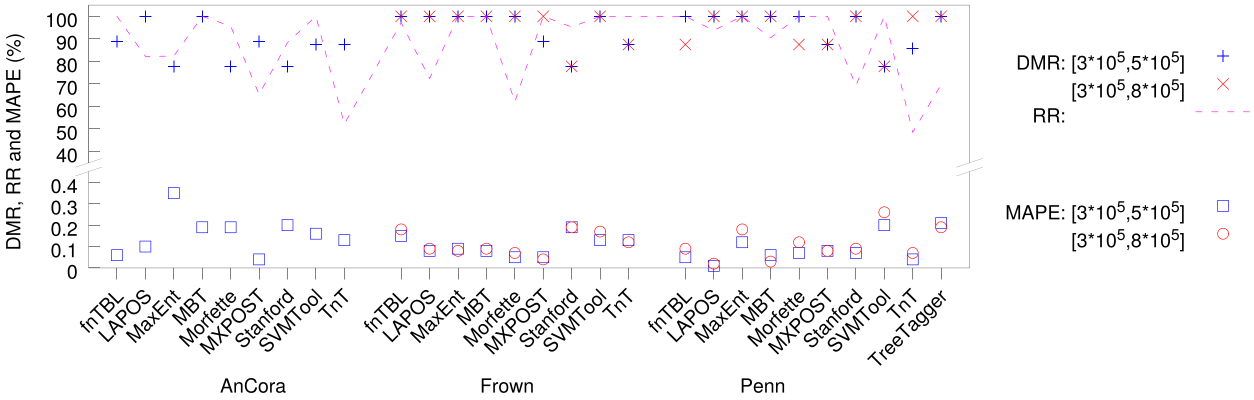

Fig. 8. MAPES and RRS for runs with anchors. DMRS when excluding crossing learning curves along the control sequences.

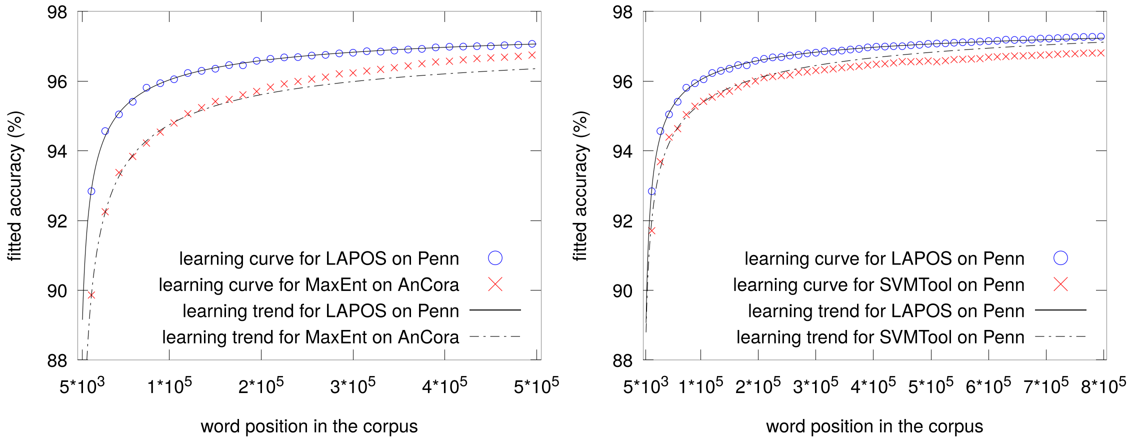

Fig. 9. Learning trends, with anchors, for the best and worst MAPES.

the corpus used to generate the observations. Out of the total of MAPE values for runs in the collection, 53.57% (resp. 57.89%) of them are in the interval [0,0.10], increasing to 92.86% (resp. 94.74%) in the interval [0,0.20].

While the results are promising for both collections $\mathcal{F}$ and $\mathcal{G}$, which leads us to argue for the goodness of the proposal on the quantitative plane, we can observe some variations by studying pairs of homologous runs. We then find that MAPEs are lower in $\mathcal{F}$, when anchors are discarded, for 67.86% (resp 73.68%) of those pairs in the interval $\left[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket\right]$ (resp. $\left[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket\right]$), and higher only in 10.71% (resp 15.79%) of them. This imbalance is due to the delay in convergence introduced by anchoring, a mechanism that increases robustness at the price of a possible slowdown of convergence. Nevertheless, the effects of this delay seem minor, given that the average difference between the MAPEs of homologous runs is 0.02 (resp. 0.03) in the interval $\left[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket\right]$ (resp. $\left[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket\right]$). At the same time, a common feature for runs in both $\mathcal{F}$ and $\mathcal{G}$ is that, when the length of the corpora allows us to overlay the control sequences, the MAPEs tend not to increase significantly between them, illustrating the reliability of calculations.

6.2.3. *The qualitative study*

Our reference performance metrics here are the DMR and the RER, depending on whether the testing scenario considered involves disjoint learning curves or not. Unlike MAPE values, we are now interested in those close to 100, the maximum possible for both metrics.

6.2.3.1. Runs involving disjoint learning curves. As regards the collection $\mathcal{F}$ of runs based on learning traces without anchors, DMRs in Fig. 6 are taken from the data in Table 1 and range from 77.78 to 100 in both control sequences. Moreover, 85.71% (resp. 89.47%) of these values belong to the interval $[87.50, 100]$ for the control sequence in $\left[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket\right]$ (resp. $\left[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket\right]$). Similar results, shown in Fig. 8 from the data compiled in Table 2, were obtained on the collection $\mathcal{G}$ of runs with anchoring learning traces. This time, DMR values range between 77.78 and 100 in both control sequences, with 78.57% (resp. 89.47%) of them in the interval $[87.50, 100]$ for the control levels in $\left[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket\right]$ (resp. $\left[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket\right]$). The DMRs prove, therefore, to be reasonably high in all the case studies. With regard to the overlap of control sequences, the values remain stable.

Finally, the delay in convergence observed in runs of $\mathcal{G}$ with respect to their homologs in $\mathcal{F}$ has a limited impact on DMRs, supporting again the use of anchoring as practical mechanism to improve the robustness. Thus, only a 14.29% (resp. 5.26%) of homologous DMRs are higher for runs without anchors in the interval $\left[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket\right]$ (resp. $\left[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket\right]$), while the remaining 85.71% (resp. 94.73%) have identical values. That reduces the average difference between homologous DMRs to 2.08 (resp. 0.66).

6.2.3.2. Runs involving intersecting learning curves. In the case of the collection $\mathcal{F}$ of runs based on learning traces without anchors, RERs for pairs in Fig. 10 are taken from the data in Table 3 and range from 31.71 (resp. 37.62) to 97.56 (resp. 99.01) for the control sequence in $\left[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket\right]$ (resp. $\left[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket\right]$). In addition, while 66.67% (resp. 33.33%) of those values can be found in the interval $[87.80, 100]$, 83.33% (resp. 66.67%) of them are in $[77.23, 100]$. Results are similar for pairs of runs with anchoring learning traces in the collection $\mathcal{G}$, which are shown in Fig. 11 from the data compiled in Table 4. Maximum and minimum RER values are the same as those obtained without anchors in both control sequences. Moreover, 66.67% (resp. 33.33%) of them can be found in the

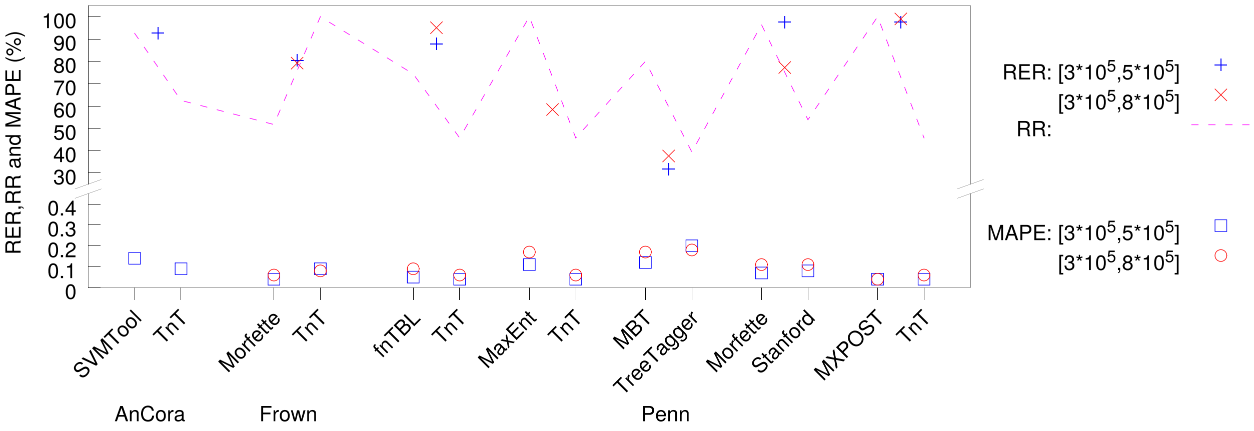

Fig. 10. MAPEs, RRs and RERs for pairs of runs, without anchors, involving crossing learning curves along the control sequences.

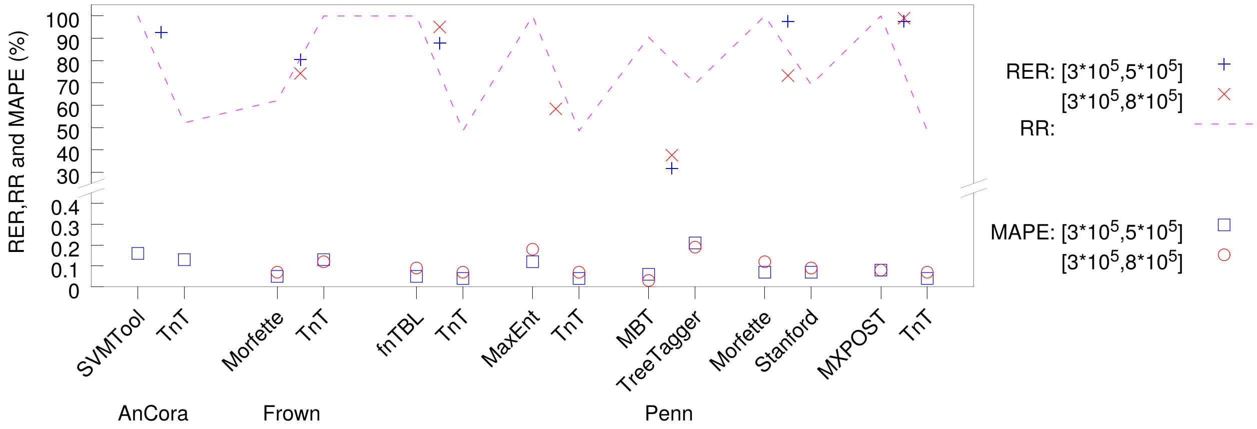

Fig. 11. MAPEs, RRs and RERs for pairs of runs, with anchors, involving crossing learning curves along the control sequences.

interval $[87.80, 100]$, climbing to 83.33% (resp. 66.67%) for the interval $[73.27, 100]$, for the control sequence in $[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket]$ (resp. $[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket]$). When the corpora cover both control sequences, there is no significant difference between their corresponding RERs, confirming the reliability of the calculations, which are also reasonably high for RER in all the case studies. In short, we can also argue the goodness of the proposal on the qualitative plane.

Following the previous discussion for MAPE and DMR, the impact of using anchoring also remains limited for RER, allowing us to definitively state its viability as an instrument for bettering robustness. To prove this, we consider pairs of RERs, one computed for two runs in $\mathcal{F}$ with intersecting learning traces, and the other for their homologs in $\mathcal{G}$. In all cases (resp. 66.67% of them) the RERs are identical for the control sequence in $[\llbracket 3*10^5 \rrbracket, \llbracket 5*10^5 \rrbracket]$ (resp. $[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket]$). The remaining 33.33% of pairs measured in $[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket]$ show higher RER values for runs in $\mathcal{F}$, namely without anchors, resulting in an average difference between RERs in the same pair of 1.48 in $[\llbracket 3*10^5 \rrbracket, \llbracket 8*10^5 \rrbracket]$.

6.2.4. *The study of robustness*

The reference metric is now RR, and we are interested in values as close as possible to 100, the maximum one. Regarding the collection of runs $\mathcal{F}$, when anchors are discarded, the results are shown in both Figs. 6 and 10 from the data compiled in Table 1. While RR values range from 39.39 to 100, the latter is only reached in 28.57% of the runs. This percentage rises to 39.29% for RRs in the interval [90,100], reaching 67.86% in [80,100]. Focusing now on the collection of runs $\mathcal{G}$, with anchoring learning traces, the results are shown in Figs. 8 and 11 from the data compiled in Table 2. Minimum and maximum RRs are, respectively, 48.57 and 100, with 46.43% of the runs achieving the maximum value. This proportion reaches 64.29% in the interval [90,100], and 75% in [80,100]. With respect to the behavior of RRs in pairs of homologous runs, higher values are found when using anchoring in 57.14% of those pairs, and lower ones only in 14.29% of them. The average difference between RR values for homologous runs is 9.43%. These results illustrate not only the good overall behavior of the prediction model against variations in its working hypotheses, but also the positive role that the use of anchors plays in this regard.

7. Conclusions

Our proposal arises as a response to the challenge of reducing both the training effort and the need for linguistic resources, in the generation of learning-based POS tagging systems in the NLP domain. In this context, we introduce a prediction strategy for learning curves, which we can exploit to deal with a variety of practical uses beyond the mere estimation for the final value of the accuracy associated to a tagger. Focusing on the most representative ones, it is possible to estimate the extra accuracy increase between any two levels in the corpus, which is helpful for evaluating the training effort needed to attain a certain measurement performance. Comparing the latter also becomes realistic at all training levels, providing us with a useful instrument for choosing the most efficient tagger in each case. Finally, accuracy prediction below a certain degree of convergence fixed by the user can be guaranteed, which gives us the possibility of evaluating the adequacy of tagger configuration on the basis of a fraction of its generation process. Altogether, these facilities involve both quantitative and qualitative aspects, forming a powerful tool for reducing training costs in tagger construction.

Formally, we have developed a technique whose generality permits it to be applied in a much wider context. Based on a functional analysis, we extend the classic discrete calculation of accuracy in ML to a continuous domain. The proposal is modeled as the uniform convergence of a sequence of learning trends which iteratively approximates the learning curve. Since the limit computed is a continuous curve, it is guaranteed to be free of gaps, breaks and holes. This allows us to make the predictions without disruptions due to instantaneous jumps and over the entire training data base, while ensuring their regularity. The correctness of the algorithm has been proven, including a proximity criterion, with respect to our working hypotheses. This permits us, once a point in the process called prediction level has been passed, to identify the iteration from which the estimates are below a convergence threshold fixed by the user.

Regarding the robustness and given that the monotony of the asymptotic backbone is at the basis of the correctness, our goal is to reduce the fluctuations in that sequence, the anchoring of the learning trends being the way to achieve this. With the aim of maximizing the efficiency of this mechanism, the user can fix a verticality threshold for determining the working level from which it is applied. While this is not a mandatory procedure, its implementation enables us to limit the use of anchors to that part of the process where the slopes are slight enough, avoiding an unnecessary deceleration of the convergence.

In practice, the experimental results in the NLP field of POS tagging corroborate our expectations on a wide range of particular cases for a representative sampling of taggers and corpora on both English and Spanish. This opens the door to new applications in ML, particularly in the NLP domain, this being the case in a variety of well known tasks, such as MT, text classification, parsing or any other kind of activity requiring linguistic annotation. All these are new fields of application we plan to explore in a future work.

Acknowledgments

We thank Prof. Christian Mair of the University of Freiburg, who provided us with a copy of the tagged versions of the Brown family of corpora, including the Freiburg-Brown corpus of American English. This research has been partially funded by the Spanish Ministry of Economy and Competitiveness through project FFI2014-51978-C2-1-R.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: 2nd International Symposium on Information Theory. Budapest, pp. 267–281.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Apostol, T.M., 2000. Mathematical Analysis. Narosa Book Distributors Pvt Ltd, New Delhi.

Attenberg, J., Provost, F., 2011. Inactive learning? Difficulties employing active learning in practice. ACM SIGKDD Explorations Newsletter 12 (2), 36–41.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36 (1–2), 105–139.

Becker, M., Osborne, M., 2005. A two-stage method for active learning of statistical grammars. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. Edinburgh, pp. 991–996.

Bertoldi, N., Cettolo, M., Federico, M., Buck, C., 2012. Evaluating the learning curve of domain adaptive statistical machine translation systems. In: Proceedings of the 7th Workshop on Statistical Machine Translation. Montreal, pp. 433–441.

Biemann, C., 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Sydney, pp. 7–12.

Birch, A., Osborne, M., Koehn, P., 2008. Predicting success in machine translation. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, pp. 745–754.

Bloodgood, M., Vijay-Shanker, K., 2009. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In: Proceedings of the 13th Conference on Computational Natural Language Learning. Boulder, pp. 39–47.

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., 1987. Occam's razor. Inf. Process. Lett. 24 (6), 377–380.

Branch, M.A., Coleman, T.F., Li, Y., 1999. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM J. Sci. Comput. 21 (1), 1–23.

Brants, T., 2000. TnT: A statistical part-of-speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing. Seattle, pp. 224–231.

Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 1996b. Bias, Variance, and Arcing Classifiers. Tech. Rep. 460. Statistics Department, University of California.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput. Linguist. 21 (4), 543–565.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.

Chan, Y.S., Ng, H.T., 2007. Domain adaptation with active learning for word sense disambiguation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, pp. 49–56.

Charniak, E., 2000. A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference. Seattle, pp. 132–139.

Chen, J., Schein, A., Ungar, L., Palmer, M., 2006. An empirical study of the behavior of active learning for word sense disambiguation. In: Proceedings of the 2006 Annual Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies. New York, pp. 120–127.

Chen, S.F., Mangu, L., Ramabhadran, B., Sarikaya, R., Sethy, A., 2009. Scaling shrinkage-based language models. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding. Merano, pp. 299–304.

Chrupala, G., Dinu, G., van Genabith, J., 2008. Learning morphology with Morfette. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, pp. 2362–2367.

Clark, A., Fox, C., Lappin, S., 2010. The Handbook of Computational Linguistics and Natural Language Processing. John Wiley & Sons, Hoboken.

Cohn, D., Atlas, L., Ladner, R., 1994. Improving generalization with active learning. Mach. Learn. 15 (2), 201–221.

Collins, M., 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (Vol. 10). Philadelphia, pp. 1–8.

Culotta, A., McCallum, A., 2005. Reducing labeling effort for structured prediction tasks. In: Proceedings of the 20th National Conference on Artificial Intelligence (Vol. 2). Pittsburgh, pp. 746–751.

Daelemans, W., Zavrel, J., Berck, P., Gillis, S., 1996. MBT: A memory–based part-of-speech tagger generator. In: Proceedings of the 4th Workshop on Very Large Corpora. Copenhagen, pp. 14–27.

Dagan, I., Engelson, S.P., 1995. Committee-based sampling for training probabilistic classifiers. In: Proceedings of the 12th International Conference on Machine Learning. Tahoe City, pp. 150–157.

DeRose, S.J., 1988. Grammatical category disambiguation by statistical optimization. Comput. Linguist. 14 (1), 31–39.

Domingo, C., Gavaldà, R., Watanabe, O., 2002. Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Min. Knowl. Discov. 6 (2), 131–152.

Floyd, S., Warmuth, M., 1995. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. Mach. Learn. 21 (3), 269–304.

Francis, W.N., Kučera, H., 1967. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Brown University, Providence.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning. Bari, pp. 148–156.

Frey, L., Fischer, D., 1999. Modeling decision tree performance with the power law. In: Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics. Fort Lauderdale, pp. 59–65.

García-Pedrajas, N., De Haro-García, A., 2014. Boosting instance selection algorithms. Knowl. Based Syst. 67, 342–360.

Giesbrecht, E., Evert, S., 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: Proceedings of the 5th Web as Corpus Workshop. San Sebastian, pp. 27–35.

Giménez, J., Márquez, L., 2004. SVMTool: A general POS tagger generator based on support vector machines. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, pp. 43–46.

Goldwater, S., Griffiths, T., 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, pp. 744–751.

Gu, B., Hu, F., Liu, H., 2001. Modelling classification performance for large data sets. In: Proceedings of the 2nd International Conference on Advances in Web-Age Information Management. Xi'an, pp. 317–328.

Haertel, R., Ringger, E., Seppi, K., Carroll, J., McClanahan, P., 2008. Assessing the costs of sampling methods in active learning for annotation. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Columbus, pp. 65–68.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M.A., Màrquez, L., et al., 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Proceedings of the 13th Conference on Computational Natural Language Learning: Shared Task. Boulder, pp. 1–18.

Hinrichs, L., Smith, N., Waibel, B., 2010. Manual of information for the part-of-speech-tagged, post-edited 'Brown' corpora. ICAME J. 34, 189–233.

Howard, R.A., 1966. Decision analysis: Applied decision theory. In: Proceedings of the 4th International Conference on Operational Research. Cambridge, pp. 55–71.

Hulden, M., Francom, J., 2012. Boosting statistical tagger accuracy with simple rule-based grammars. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, pp. 2114–2117.

Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.

John, G., Langley, P., 1996. Static versus dynamic sampling for data mining. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, pp. 367–370.

Kapoor, A., Greiner, R., 2005. Learning and classifying under hard budgets. In: Machine Learning: ECML 2005. Springer-Verlag, pp. 170–181.

Koehn, P., Och, F.J., Marcu, D., 2003. Statistical phrase-based translation. In: Proceedings of the 2003 Annual Conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology (Vol. 1). Edmonton, pp. 48–54.

Kolachina, P., Cancedda, N., Dymetman, M., Venkatapathy, S., 2012. Prediction of learning curves in machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers (Vol. 1). Jeju Island, pp. 22–30.

Langford, J., 2005. Tutorial on practical prediction theory for classification. J. Mach. Learn. Res. 6, 273–306.

Last, M., 2009. Improving data mining utility with projective sampling. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, pp. 487–496.

Laws, F., Schütze, H., 2008. Stopping criteria for active learning of named entity recognition. In: Proceedings of the 22nd International Conference on Computational Linguistics (Vol. 1). Manchester, pp. 465–472.

Lebreton, J.-D., Burnham, K.P., Clobert, J., Anderson, D.R., 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecol. Monogr. 62 (1), 67–118.

Leech, G., 2009. Change in Contemporary English: A Grammatical Study. Studies in English Language. Cambridge University Press, Cambridge.

Leite, R., Brazdil, P., 2007. An iterative process for building learning curves and predicting relative performance of classifiers. In: Proceedings of the Artificial Intelligence 13th Portuguese Conference on Progress in Artificial Intelligence. Guimarães, pp. 87–98.

Leung, K.T., Parker, D.S., 2003. Empirical comparisons of various voting methods in bagging. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, pp. 595–600.

Lewis, D.D., Gale, W.A., 1994. A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, pp. 3–12.

Li, S., Graça, J.V., Taskar, B., 2012. Wiki-ly supervised part-of-speech tagging. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, pp. 1389–1398.

Liere, R., Tadepalli, P., 1997. Active learning with committees for text categorization. In: Proceedings of the 14th National Conference on Artificial Intelligence. Providence, pp. 591–596.

Mair, C., 2006. Twentieth-Century English: History, Variation and Standardization. Studies in English Language. Cambridge University Press, Cambridge.

Mair, C., Leech, G., 2007. The Freiburg-Brown corpus (‘Frown’) (POS-tagged version).

Marcus, M.P., Santorini, B., Marcinkiewicz, M.A., Taylor, A., 1999. Treebank-3 LDC99T42. Web Download file. Linguistic Data Consortium, Philadelphia.

McAllester, D.A., 1999. PAC-Bayesian model averaging. In: Proceedings of the 12th Annual Conference on Computational Learning Theory. Santa Cruz, pp. 164–170.

McCallum, A., Nigam, K., 1998. Employing EM and pool-based active learning for text classification. In: Proceedings of the 15th International Conference on Machine Learning. Madison, pp. 350–358.

Meek, C., Thiesson, B., Heckerman, D., 2002. The learning-curve sampling method applied to model-based clustering. J. Mach. Learn. Res. 2, 397–418.

Merialdo, B., 1994. Tagging English text with a probabilistic model. Comput. Linguist. 20 (2), 155–171.

Monachini, M., Calzolari, N., 1996. EAGLES Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages. Tech. rep., Centre National de la Recherche Scientifique.

Mukund, S., Ghosh, D., Srihari, R.K., 2010. Using cross-lingual projections to generate semantic role labeled corpus for Urdu: a resource poor language. In: Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, pp. 797–805.

Neubig, G., Nakata, Y., Mori, S., 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Vol. 2). Portland, pp. 529–533.

Ngai, G., Florian, R., 2001. Transformation-based learning in the fast lane. In: Proceedings of the 2nd Meeting of the North American chapter of the Association for Computational Linguistics on Language Technologies. Pittsburgh, pp. 1–8.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., et al., 2007. The CoNLL-2007 shared task on dependency parsing. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, pp. 915–932.

Petrov, S., Barrett, L., Thibaux, R., Klein, D., 2006. Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, pp. 433–440.

Popel, M., Mareček, D., Štěpánek, J., Zeman, D., Žabokrtský, Z., 2013. Coordination structures in dependency treebanks. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 1). Sofia, pp. 517–527.

Provost, F., Jensen, D., Oates, T., 1999. Efficient progressive sampling. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, pp. 23–32.

Ratnaparkhi, A., 1996. A maximum entropy model for part-of-speech tagging. In: Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing. Philadelphia, pp. 133–142.

Ravi, S., Knight, K., 2009. Minimized models for unsupervised part-of-speech tagging. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (Vol. 1). Suntec, pp. 504–512.

Recasens, M., Màrquez, L., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., et al., 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, pp. 1–8.

Ringger, E., McClanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., et al., 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In: Proceedings of the Linguistic Annotation Workshop. Prague, pp. 101–108.

Roy, N., McCallum, A., 2001. Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of the 18th International Conference on Machine Learning. Williamstown, pp. 441–448.

Sarikaya, R., Çelikyilmaz, A., Deoras, A., Jeong, M., 2014. Shrinkage based features for slot tagging with conditional random fields. In: 15th Annual Conference of the International Speech Communication Association. Singapore, pp. 268–272.

Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5 (2), 197–227.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester, pp. 44–49.

Schütze, H., Velipasaoglu, E., Pedersen, J.O., 2006. Performance thresholding in practical text classification. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. Arlington, pp. 662–671.

Seung, H.S., Opper, M., Sompolinsky, H., 1992. Query by committee. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, pp. 287–294.

Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.-L., 2004. Multi-criteria-based active learning for named entity recognition. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, pp. 589–596.

Sheng, V.S., Ling, C.X., 2007. Partial example acquisition in cost-sensitive learning. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, pp. 638–646.

Song, H.-J., Son, J.-W., Noh, T.-G., Park, S.-B., Lee, S.-J., 2012. A cost sensitive part-of-speech tagging: Differentiating serious errors from minor errors. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers (Vol. 1). Jeju Island, pp. 1025–1034.

Søgaard, A., 2010. Simple semi-supervised training of part-of-speech taggers. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Short Papers. Uppsala, pp. 205–208.

Spoustová, D., Hajič, J., Raab, J., Spousta, M., 2009. Semi-supervised training for the averaged perceptron POS tagger. In: Proceedings of the 12th Conference of the European chapter of the Association for Computational Linguistics. Athens, pp. 763–771.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J., 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In: Proceedings of the 12th Conference on Computational Natural Language Learning. Manchester, pp. 159–177.

Tang, M., Luo, X., Roukos, S., 2002. Active learning for statistical natural language parsing. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, pp. 120–127.

Taulé, M., Martí, M.A., Recasens, M., 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, pp. 96–101.

Thompson, C.A., Califf, M.E., Mooney, R.J., 1999. Active learning for natural language parsing and information extraction. In: Proceedings of the 16th International Conference on Machine Learning. Bled, pp. 406–414.

Tomanek, K., Hahn, U., 2008. Approximating learning curves for active-learning-driven annotation. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, pp. 1319–1324.

Tomanek, K., Wermter, J., Hahn, U., 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, pp. 486–495.

Tong, S., Koller, D., 2002. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2, 45–66.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Annual Conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology (Vol. 1). Edmonton, pp. 173–180.

Tsuruoka, Y., Miyao, Y., Kazama, J., 2011. Learning with lookahead: Can history-based models rival globally optimized models? In: Proceedings of the 15th Conference on Computational Natural Language Learning. Portland, pp. 238–246.

Turchi, M., De Bie, T., Cristianini, N., 2008. Learning performance of a machine translation system: A statistical and computational analysis. In: Proceedings of the 3rd Workshop on Statistical Machine Translation. Columbus, pp. 35–43.

Turney, P.D., 2000. Types of cost in inductive concept learning. In: Workshop on Cost-Sensitive Learning at the 17th International Conference on Machine Learning. Stanford, pp. 15–21.

van Halteren, H., 1999. Performance of taggers. In: Syntactic Wordclass Tagging. Kluwer Academic Pub., Hingham, pp. 81–94.

Valiant, L.G., 1984. A theory of the learnable. Commun. ACM 27 (11), 1134–1142.

Vandome, P., 1963. Econometric forecasting for the United Kingdom. Oxf. Bull. Econ. Stat. 25, 239–281.

Vlachos, A., 2008. A stopping criterion for active learning. Comput. Speech Lang. 22 (3), 295–312.

Weiss, G., Tian, Y., 2008. Maximizing classifier utility when there are data acquisition and modeling costs. Data Min. Knowl. Discov. 17 (2), 253–282.

Weiss, G.M., Provost, F., 2003. Learning when training data are costly: the effect of class distribution on tree induction. J. Artif. Intell. Res. 19 (1), 315–354.

Yuret, D., Han, A., Turgut, Z., 2010. SemEval-2010 Task 12: Parser evaluation using textual entailments. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Los Angeles, pp. 51–56.

Zhu, J., Hovy, E., 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, pp. 783–790.

Zhu, J., Ma, M., 2012. Uncertainty-based active learning with instability estimation for text classification. ACM Trans. Speech Lang. Process. 8 (4), 5:1–5:21.