Article

Surfing the Modeling of POS Taggers in Low-Resource Scenarios

Manuel Vilares Ferro [1,*,†] **, Víctor M. Darriba Bilbao** [1,†] **, Francisco J. Ribadas Pena** [1,†] **and Jorge Graña Gil** [2,†]

1. Department of Computer Science, University of Vigo, Edificio Politécnico, As Lagoas s/n,
   32004 Ourense, Spain
2. Department of Computer Science, Faculty of Informatics, University of A Coruña, 15071 A Coruña, Spain
* Correspondence: vilares@uvigo.es
† These authors contributed equally to this work.

Abstract: The recent trend toward the application of deep structured techniques has revealed the limits of huge models in natural language processing. This has reawakened the interest in traditional machine learning algorithms, which have proved still to be competitive in certain contexts, particularly in low-resource settings. In parallel, model selection has become an essential task to boost performance at reasonable cost, even more so when we talk about processes involving domains where the training and/or computational resources are scarce. Against this backdrop, we evaluate the early estimation of learning curves as a practical mechanism for selecting the most appropriate model in scenarios characterized by the use of non-deep learners in resource-lean settings. On the basis of a formal approximation model previously evaluated under conditions of wide availability of training and validation resources, we study the reliability of such an approach in a different and much more demanding operational environment. Using as a case study the generation of POS taggers for Galician, a language belonging to the Western Ibero-Romance group, the experimental results are consistent with our expectations.

Keywords: learning curves; low-resource scenarios; non-deep machine learning; model selection; POS taggers; stopping criteria

MSC: 68; 68T50

1. Introduction

The application of *machine learning* (ML) techniques has significantly changed the landscapes of *natural language processing* (NLP) over the last decade, allowing some of the gaps derived from the use of rule-based approaches to be filled in, mainly its lack of flexibility and high development cost. Thus, although the state of the art makes clear that hand-crafted tools are not only easier to interpret and manipulate but also often provide better results [1–5], the high level of dependency on expert knowledge makes their implementation and subsequent maintenance costly in human terms, in addition to hindering their applicability to different languages [2,5–8]. Combined with the surge in computational power, the possibility of accessing massive amounts of data and the decline in the cost of disk storage, this has decisively contributed to the growing popularity of ML algorithms as the basis for classification [9] and clustering [10] models in a variety of tasks. This includes entity detection [11], information retrieval [12], language identification [13], machine translation [14], question answering [15], semantic role labeling [16], sentiment analysis [17] and text classification [18], among others.

At this juncture, although recent proposals based on *deep learning* (DL) have outperformed the traditional ML methods on a variety of operating fronts, the approach has also shown its limitations. Specifically, there are two main reasons for the popularity and the arguable superiority of the DL solutions: *end-to-end training* and *feature learning*—while end-to-end training allows the model to learn all the steps between the initial input phase

and the final output result, feature learning offers representability to effectively encode the information in the data. However, the latter translates into the inability to handle directly symbols [19]. This implies that data must first be converted to vector representations for input into the model and then do just the reverse with its output, which leads to a complex interpretability of models. Other well-known challenges are the lack of theoretical foundation [20], the difficulty in dealing with the long tail [21,22], the ineffectiveness of inference and decision making [23], and the requirement of large amounts of data and powerful computing resources that may not be available. This complex picture looks even worse in the NLP domain [24], particularly when it comes to dealing with *low-resource scenarios*—i.e., languages, domains or tasks lacking large corpora and/or manually crafted linguistic resources sufficient for building NLP applications— [25]. On the one hand, feature-based techniques lead to an imperfect use of the linguistic knowledge. On the other, the scarcity of training data is not only problematic *per se*, but also because of its impact on the rest of trials. So, generating high-quality vector representations remains a challenge [26] and the imbalance in the training samples that start the long tail and bias phenomena is more likely, together with the proneness to overfitting of DL models [27], which can result in poor predictive power, thereby compromising both inference and decision making.

This has led to a renewal of interest in reviewing the role of DL techniques vs. traditional ML ones in the development of NLP applications [3,24,28–30], particularly in low-resource scenarios [25]. Special attention has been given here to sequence labeling tasks [30]. These encompass a class of NLP problems that involve the assignment of a categorical label to each member of a sequence of observed values, and whose output facilitates downstream applications, such as parsing or semantic analysis, so errors at this stage can lower their performance [31]. Among the most important, we can highlight named entity recognition [7,32], multi-word expression identification [29], and morphological [26] and POS tagging [2,33–36]. It is precisely in this framework, the generation of POS taggers for low-resource scenarios by means of non-deep ML, that we propose the study of model selection based on the early estimation of learning curves. With that in mind, we first overview the state of the art and our contribution in Section 2. Next, Section 3 briefly reports on the theoretical basis supporting our research. In Section 4, we introduce the testing frame for the experiments described in Section 5 and later discussed in Section 6. Finally, Section 7 presents our conclusions and thoughts for future work.

2. Related Work and Contribution

Model selection based on the estimation of learning curves has been the subject of ongoing research over recent decades, inspired by the idea that the loss of predictive power and of training are correlated [37]. In the scope of NLP, these techniques have been applied to the most commonly researched areas, such as, for example, *machine translation*. Specifically, they have been used here for assessing the quality systems [38,39], optimizing parameter setting [40], estimating how much training data is required to achieve a certain degree of accuracy [41] or evaluating the impact of a concrete set of distortion factors on the performance [42]. Their popularity is growing especially in the field of *active learning*—those iterative ML strategies that interact with the environment in each cycle, selecting for annotation the instances which are harder to identify—(AL) [43], where we can refer to applications for information extraction [44,45], parsing [46,47] and text classification [48–51]. The same is true for POS tagging [52–54] and closely related tasks, such as named entity recognition [55–57] or word sense disambiguation [58–60], always with the aim of reducing the annotation effort. Since AL prioritizes the data to be labeled in order to maximize the impact for training a supervised model, it performs better than other ML strategies with substantially fewer resources. This justifies the interest in it as an underlying learning guideline to deal with low-resource scenarios [61–64] and specifically in the area of POS tagging [54,65–69].

Focusing on the early estimation of learning curves in AL, we can distinguish between *functional* [55,70] and *probabilistic* [71–73] proposals, depending on the nature of the halting

condition used to determine the end of the training process from the information generated in each cycle. As a basic difference, functional strategies not only permit the calculation of relative and absolute error (respectively, convergence) thresholds —an error (respectively, convergence) threshold measures the difference between the real and the estimated learning curves at a finite (respectively, infinite) approximation time, while the absolute or relative character is applicable to any type of estimation, referring in the first case to the strict difference between the values compared, and to that existing between values calculated during the prediction process in the second one— [74], but they are also simpler and more robust than techniques based on probabilistic ones. In particular, by replacing single observations with distributions, we introduce elements of randomness, and thus uncertainty. That way, to establish how much data is necessary to reliably build such distributions is no easy matter, and the same applies to rare event handling. Being related to the definition of a sampling strategy on a sufficiently wide range of observations, this question should be preventable or better dealt with in a functional frame [75], even more so when the scarcity of training resources make it difficult to apply probabilistic criteria.

In such a context, we face the evaluation of POS tagging models by early estimation of learning curves when working in resource-scarce settings, to the best of our knowledge, a yet unexplored terrain. Leading on from this and looking for an operational solution, we focus on AL scenarios, turning our attention to a functional view of the issue. With a view to exploring the practicality and potential of the approach, we address it in a setting used previously to demonstrate its effectiveness when the availability of resources for learning is not a problem. That way, we take up both the formal prediction concept and the testing frame introduced in [70], which also allow us to contrast the level of efficiency to be expected when the conditions for training and validation of the generated models are much more restrictive.

3. The Formal Framework

Below is a brief review of the theoretical basis underlying our work, taken from [70]. From now on, we denote the real numbers by $\mathbb{R}$ and the natural ones by $\mathbb{N}$, assuming that $0 \notin \mathbb{N}$. The order in $\mathbb{N}$ is also extended to $\overline{\mathbb{N}} := \mathbb{N} \cup \{\infty, \boldsymbol{\infty}\}$, in such a way that $\boldsymbol{\infty} > \infty > i > 0$, $\forall i \in \mathbb{N}$. Assuming that a learning curve is a plot of model learning performance over experience, we focus on accuracy as a measure of that performance.

3.1. The Notational Support

We start with a sequence of observations calculated from cases incrementally taken from a training data base, and organized around the concept of *learning scheme* [70].

Definition 1 (Learning scheme). *Let $\mathcal{D}$ be a training data base, $\mathcal{K} \subsetneq \mathcal{D}$ a set of initial items from $\mathcal{D}$, and $\sigma : \mathbb{N} \to \mathbb{N}$ a function. We define a* learning scheme *for $\mathcal{D}$ with* kernel $\mathcal{K}$ and step σ, as a *triple $\mathcal{D}_\sigma^{\mathcal{K}} = [\mathcal{K}, \sigma, \{\mathcal{D}_i\}_{i \in \mathbb{N}}]$ such that $\{\mathcal{D}_i\}_{i \in \mathbb{N}}$ is a cover of $\mathcal{D}$ verifying*

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

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

That relates a level i with the position $\|\mathcal{D}_i\|$ in the training database, determining the sequence of observations $\{[x_i, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_i)], \ x_i := \|\mathcal{D}_i\|\}_{i \in \mathbb{N}}$, where $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_i)$ is the accuracy achieved on such an instance by the learner. Thus, a level determines an iteration in the adaptive sampling, whose learning curve is $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}]$, whilst $\mathcal{K}$ delimits a portion of $\mathcal{D}$ we believe to be enough to initiate consistent evaluations of the training. For its part, σ identifies the sampling scheduling.

In order to obtain a reliable assessment, the weak predictor generated at each learning cycle is extrapolated according to an *accuracy pattern* [70], which allows to formally compile a set of properties, giving stability to the estimates and widely accepted as *working hypotheses* by the state of the art in model evaluation [72,76–80].

Definition 2 (Accuracy pattern). *Let $C_{(0,\infty)}^\infty$ be the C-infinity functions in $\mathbb{R}^+$, we say that $\pi : \mathbb{R}^{+^n} \to C_{(0,\infty)}^\infty$ is an accuracy pattern iff $\pi(a_1, \ldots, a_n)$ is positive definite, upper bounded, concave and strictly increasing.*

As a running accuracy pattern, we select the *power law family* $\pi(a, b, c)(x) := -a * x^{-b} + c$. Its use is illustrated in the right-most diagram of Figure 1 to fit the learning curve represented on the left-hand side for the SVMTool tagger [81] on the Xiada corpus of Galician [82], with the values $a = 204.570017$, $b = 0.307277$ and $c = 99.226727$ provided by the *trust region method* [83]. Returning to the review of our notational support, we now adapt these calculation elements to an iterative dynamics through the concept of *learning trend* [70].
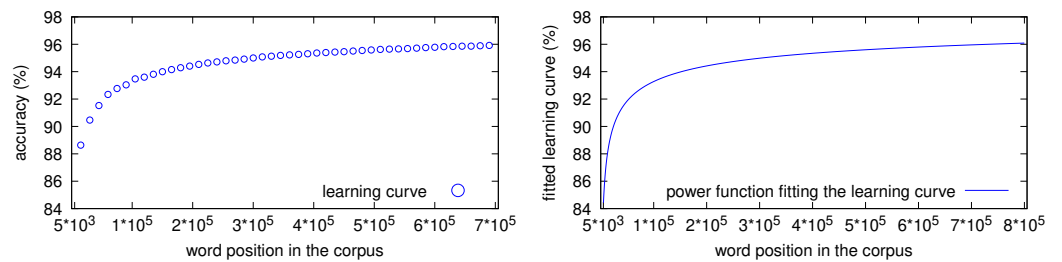

Figure 1. Learning curve for SVMTool on Xiada, and an accuracy pattern fitting it.

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

The minimum level ℓ for a learning trend $\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ is 3, because we need at least three points to generate a curve. Its value $\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}](x_i)$ is our prediction of the accuracy for a case x_i, using a model generated from the first ℓ cycles of the learner. Accordingly, the asymptotic term α_ℓ is interpretable as the estimate for the highest accuracy attainable. Continuing with the tagger SVMTool and the corpus Xiada, Figure 2 shows a portion of the learning trace with kernel and uniform step function 5×10^3, including a zoom view.
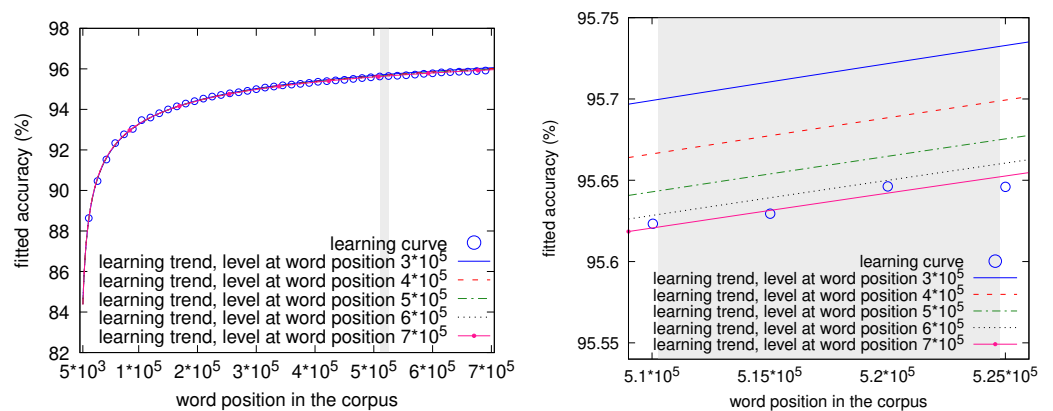

Figure 2. Learning trace for SVMTool on Xiada, with details in zoom.

3.2. Correctness

Assuming our working hypotheses, the correctness of the proposal—i.e., the existence and effective approximation of a learning curve $\mathcal{A}_\infty[\mathcal{D}]$ from a subset of its observations

compiled in a learning scheme $\mathcal{D}_\sigma^\mathcal{K}$—is demonstrated from the uniform convergence of the corresponding learning trace $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ [70]. Specifically, the function

$$\mathcal{A}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}] := \lim_{i \to \infty}{}^u \mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$$

exists and is positive definite, increasing, continuous and upper bounded by 100 in $(0, \infty)$. In order to estimate the quality of our approximation, a relative proximity criterion is introduced. Labeled *layered convergence*, it evaluates the contribution of each learning trend to the convergence process in a sequence which is proved to be decreasing and convergent to zero. These layers of convergence can then be interpreted as a reliable reference to fix error (respectively, convergence) thresholds.

3.3. Robustness

Robustness is studied from a set of *testing hypotheses*, which assume that learning curves are positive definite and upper bounded, albeit only quasi-strictly increasing and concave. An observation is then categorized according to its position with respect to the *working level* (WLevel), i.e., the cycle after which irregularities would not impact the correctness. Considering that the learner stabilizes as the training advances and that the monotonicity of the asymptotic backbone is at the basis of any halting condition, it is identified as the level providing the first slope fluctuation below a given ceiling in such a backbone and, once passed, the *prediction level* (PLevel) marks the beginning of learning trends which could feasibly predict the learning curve, namely, not exceeding its maximum (100) [70]. Based on this, WLevel is calculated as the lowest level for which the normalized absolute value of the slope of the line joining two consecutive points on the asymptotic backbone is less than a verticality threshold ν, which is corrected by applying a coefficient $1/\varsigma$ for avoiding having to deal with to both infinitely large slopes and extremely small decimal fractions.

Definition 4 (Working and prediction levels). *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 $\lambda \in \mathbb{N} \cup \{0\}$. The working level (WLevel) for $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ with verticality threshold ν, slowdown ς and look-ahead λ, is the smallest $\omega(\nu, \varsigma, \lambda) \in \mathbb{N}$ verifying*

$$\frac{\sqrt[\varsigma]{\nu}}{1 - \nu} \geq \frac{|\alpha_{i+1} - \alpha_i|}{x_{i+1} - x_i}, \ x_i := \|\mathcal{D}_i\|, \ \forall i \text{ such that } \omega(\nu, \varsigma, \lambda) \leq i \leq \omega(\nu, \varsigma, \lambda) + \lambda \quad (2)$$

while the smallest $\wp(\nu, \varsigma, \lambda) \geq \omega(\nu, \varsigma, \lambda)$ with $\alpha_{\wp(\nu, \varsigma, \lambda)} \leq 100$ is the prediction level (PLevel).

Following our example, Figure 3 shows the scale of the deviations in the asymptotic backbone before and after WLevel, for $\nu = 2 \times 10^{-5}$, $\varsigma = 1$ and $\lambda = 5$. The PLevel is also displayed, proving that these two levels might not be the same.
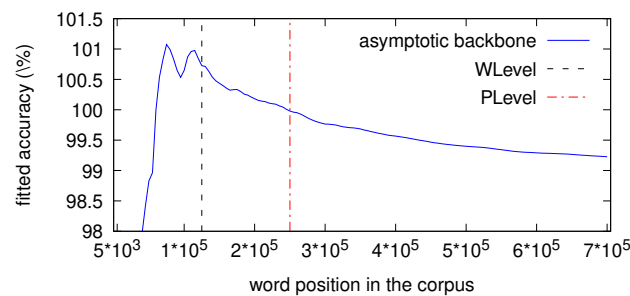

Figure 3. Working and prediction levels for SVMTool on Xiada, with details in zoom.

4. The Testing Frame

Given a training corpus $\mathcal{D}$, we want to study how far in advance and how well a learning curve $\mathcal{A}_\infty[\mathcal{D}_\sigma^\mathcal{K}]$, built from a kernel $\mathcal{K}$ and using a step function σ, can be

approximated in a low-resource scenario. To ensure the relevance of the results obtained, we standardize the conditions under which the experiments take place, following the same criteria previously considered in the study of resource-rich languages [70].

4.1. The Monitoring Structure

As an evaluation basis, we consider the *run* [70], a tuple $\mathcal{E} = [\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \wp(\nu, \varsigma, \lambda), \tau]$ characterized by a learning trace $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$, a prediction level $\wp(\nu, \varsigma, \lambda)$ and a convergence threshold τ. We then apply our study on a collection of runs $\mathcal{C} = \{\mathcal{E}_i\}_{i \in I}$, defined for a set of different learners. In order to avoid misconceptions due to the lack of uniformity in the testing frame, a common corpus $\mathcal{D}$, kernel size, accuracy pattern π, step function σ, verticality threshold ν, slowdown ς, look-ahead λ and convergence threshold τ are used.

In practice, we are interested in studying each run $\mathcal{E} = [\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \wp(\nu, \varsigma, \lambda), \tau]$ from the level in which predictions are below τ, and which we baptize *convergence level* (CLevel). So, once the PLevel is found during the computation of the trace $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$, we begin to check the layer of convergence. When it reaches the threshold τ, the trend $\mathcal{A}_{\mathrm{CLevel}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ becomes the model for the learning curve $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$, and the process of approximation is stopped.

For the runs $\mathcal{C} = \{\mathcal{E}_i\}_{i \in I}$, monitoring is applied to the learning trends $\{\mathcal{A}_{\mathrm{CLevel}_i}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]\}_{i \in I}$ on a finite common *control sequence* of levels for the training data base, which are extracted from an interval of the *prediction windows* $\{[\mathrm{CLevel}_i, \infty)\}_{i \in I}$ [70]. In each *control level*, the *accuracy* (Ac) and the corresponding *estimated accuracy* (EAc) are computed for each run using six decimal digits, though only two are represented for reasons of space and visibility.

4.2. The Performance Metrics

Our aim is both to assess the reliability of our estimates and their robustness against variations in the working hypotheses. To do so, we employ two specific kind of metrics [70].

4.2.1. Measuring the Reliability

We here differentiate two complementary viewpoints: quantitative and qualitative. In the first case, it is simply a matter of studying the closeness of the estimates and the actual learning curves, while in the second, the objective is to determine the impact of those estimates on the decision making about the performance of some models relative to others.

The Quantitative Perspective

A simple way of measuring the reliability from this viewpoint is through the *mean absolute percent error* (MAPE) [84]. 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 $\mathcal{A}_{\mathrm{CLevel}_{\mathcal{E}}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}](i)$ and the Ac from $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}_{\sigma}^{\mathcal{K}}](i)$. We can then express the MAPE as the arithmetic mean of the unsigned PE [70], as

$$\mathrm{PE}(\mathcal{E})(i) := 100 * \frac{[\mathcal{A}_{\mathrm{CLevel}_{\mathcal{E}}}^{\pi} - \mathcal{A}_{\boldsymbol{\infty}}][\mathcal{D}_{\sigma}^{\mathcal{K}}](i)}{\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}_{\sigma}^{\mathcal{K}}](i)}, \; \mathcal{E} = [\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \wp(\nu, \varsigma, \lambda), \tau], \; i \in \mathcal{S} \quad (3)$$

$$\mathrm{MAPE}(\mathcal{E})(\mathcal{S}) := \frac{100}{\|\mathcal{S}\|} * \sum_{i \in \mathcal{S}} | \, \mathrm{PE}(\mathcal{E})(i) \, | \quad (4)$$

Intuitively, the error in the estimates made over a control sequence is, on average, proportional to the MAPE, which fulfills our requirements at this point.

The Qualitative Perspective

To that end, having fixed a collection of runs $\mathcal{H}$ working on a common corpus and a control sequence $\mathcal{S}$, the reliability of one such run depends on the percentage of cases in which its estimates do not alter the relative position of its learning curve with respect to the

rest throughout $\mathcal{S}$. In this sense, our primary reference is the *reliability estimation* (RE) *of two runs* $\mathcal{E}, \tilde{\mathcal{E}} \in \mathcal{H}$ *on* $i \in \mathcal{S}$, defined [70] as

$$\mathrm{RE}(\mathcal{E}, \tilde{\mathcal{E}})(i) := \begin{cases} 1 \text{ if } [[\mathcal{A}_{\boldsymbol{\infty}} - \tilde{\mathcal{A}}_{\boldsymbol{\infty}}] * [\mathcal{A}^{\pi}_{\mathrm{CLevel}_{\mathcal{E}}} - \tilde{\mathcal{A}}^{\pi}_{\mathrm{CLevel}_{\mathcal{E}}}]][\mathcal{D}^{\mathcal{K}}_{\sigma}](i) \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

with $\mathcal{E} = [\mathcal{A}^{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}], \wp(\nu, \varsigma, \lambda), \tau]$, $\tilde{\mathcal{E}} = [\tilde{\mathcal{A}}^{\pi}[\mathcal{D}^{\tilde{\mathcal{K}}}_{\tilde{\sigma}}], \wp(\nu, \varsigma, \lambda), \tilde{\tau}]$ 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 it can be extended to the control sequence through the concept of the *reliability estimation ratio* [70].

Definition 5 (Reliability estimation ratio). *Let $\mathcal{E}$ and $\tilde{\mathcal{E}}$ 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*

$$\mathrm{RER}(\mathcal{E}, \tilde{\mathcal{E}})(\mathcal{S}) := 100 * \frac{\sum_{i \in \mathcal{S}} \mathrm{RE}(\mathcal{E}, \tilde{\mathcal{E}})(i)}{\|\mathcal{S}\|} \tag{6}$$

From this, we can calculate the percentage 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. We denote the resulting metric as *decision-making reliability* [70].

Definition 6 (Decision-making reliability). *Let $\mathcal{H} = \{\mathcal{E}_k\}_{k \in K}$ and $\mathcal{E} \notin \mathcal{H}$ 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*

$$\mathrm{DMR}(\mathcal{E}, \mathcal{H})(\mathcal{S}) := 100 * \frac{\|\mathcal{E}_k \in \mathcal{H}, \ \mathrm{RER}(\mathcal{E}, \mathcal{E}_k)(\mathcal{S}) = 100\|}{\|\mathcal{S}\|} \tag{7}$$

4.2.2. Measuring the Robustness

Since the stability of a run $\mathcal{E}$ correlates to the degree of monotony in its asymptotic backbone, we measure it as the percentage of monotonic elements in the latter through the interval $[\mathrm{WLevel}_{\mathcal{E}}, \mathrm{CLevel}_{\mathcal{E}}]$ where the approximation performs effectively. We baptize it as the *robustness rate* [70].

Definition 7 (Robustness rate). *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*

$$\mathrm{RR}(\mathcal{E}) := 100 * \frac{\|\mu\|}{\|\{\alpha_i, \ \mathrm{WLevel}_{\mathcal{E}} \leq i \leq \mathrm{CLevel}_{\mathcal{E}}\}\|} \tag{8}$$

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

The tolerance of a run to variations in the working hypotheses is therefore greater the higher its RR, thus providing a simple criterion for checking the degree of robustness on which we can count.

5. The Experiments

Within a model selection context and focused on the generation of POS taggers in low-resource scenarios, our goal is to provide evidence of the suitability of using evaluation mechanisms based on the early estimation of learning curves. It is a non-trivial challenge because, to provide that evidence, we need to study our estimates over a significant range of observations, which is in clear contradiction to the scarcity of training data.

5.1. The Linguistics Resources

In order to address the issue posed, we chose to work with a case study that meets four conditions. The first is that the language considered is really a resource-poor one, which

should guarantee that it has been outside the tuning phase of the process of developing the taggers later used in the experiments, thereby precluding any potential biases associated with the learning architecture. Second, it should have a rich morphology, thus making the training process non-trivial and therefore relevant to the test performed. Thirdly, we should have at least a training corpus of sufficient size to study the reliability of the results obtained. Finally, that corpus should provide sufficiently low levels of convergence to allow the identification of the learning processes with the generation of viable models from a small set of training data.

We then take as a case study Galician, a member of the West Iberian group of Romance languages that also includes the better-known Portuguese. It is an inflectional language with a great variety of morphological processes, particularly non-concatenative ones, derived from its Latin origin. Some of its most distinctive characteristics are as follows [85]:

- A highly complex conjugation paradigm, with 10 simple tenses including the infinitive conjugate, all of which have 6 different persons. If we add the present imperative with 2 forms, non-conjugated infinitive, gerund and participle, then 65 inflected forms are associated with each verb.
- Irregularities in both verb stems and endings. Common verbs, such as `facer` (*to do*), have up to five stems: `fac-er`, `fag-o`, `fa-s`, `fac-emos`, `fix-en`. Approximately 30% of verbs are irregular.
- Verbal forms with enclitic pronouns at the end, which can produce changes in the stem due to the presence of accents: `deu` (*gave*), `déullelo` (*he/she gave it to them*). The unstressed pronouns are usually suffixed and, moreover, they can be easily drawn together and often are contracted (`lle + o = llo`), as in the case of `váitemello buscar` (*go and fetch it for him (do it for me)*). It is also frequent to use what we call a *solidarity pronoun*, as `che` and `vos`, in order to let the listeners be participants in the action. That way, forms with up to four enclitic pronouns, such as `perdéuchellevolo` (*he had lost it to him*), are rather common.
- A highly complex gender inflection, including words with only one gender, such as `home` (*man*) and `muller` (*woman*), and words with the same form for both genders as `azul` (*blue*). Regarding words with separate forms for masculine and feminine, more than 30 variation groups are identified.
- A highly complex number inflexion, with words only being presented in singular form, such as `luns` (*monday*), and others where only the plural form is correct, such as `matemáticas` (*mathematics*). More than a dozen variation groups are identified.

This choice limits the availability of curated corpora of sufficient size to a single candidate, Xiada [82], whose latest version (2.8) includes over 747,000 entries gathered from three different sources: general and economic news articles, and short stories. With the aim of accommodating the elaborate linguistic structure previously described, the tag-set includes 460 tags, a short description of which can be found at http://corpus.cirp.gal/xiada/etiquetario/taboa (accessed on 13 July 2022).

5.2. The POS Tagging Systems

As already argued, we focus on models built from AL, selecting a broad range of proposals covering the most representative non-deep learning architectures—our reference here is the state of the art in POS tagging by the *Association for Computational Linguistics* (ACL), available at the link https://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art) (accessed on 13 July 2022)—, the same tested in [70] on resource-rich languages. This matching will allow us to establish, together with the subsequent identification of the parameters in the testing space, a valid reference based on the results obtained in that work:

- In the category of stochastic methods and representing the *hidden Márkov models* (HMMs), we choose TnT [86]. We also include the TreeTagger [87], a proposal that uses decision trees to generate the HMM, and Morfette [88], an averaged perceptron approach [89]. To illustrate the *maximum entropy models* (MEMs), we select MXPOST [90]

and OpenNLP MaxEnt [91]. Finally, the Stanford POS tagger [91] combines features of HMMs and MEMs using a *conditional Márkov model*.

- Under the heading of other approaches, we consider fnTBL [92], an update of the classic Brill tagger [93], as an example of transformation-based learning. As a memory-based method, we take the *memory-based tagger* (MBT) [94], while SVMTool [81] illustrates the behavior of *support vector machines* (SVMs).

In addition, this ensures an adequate coverage of the range of learners available in the computational domain under consideration.

5.3. The Testing Space

Following the method drawn by the choice of ML architectures discussed above, the design of the testing space is the same as the one considered in the state of the art [70] for the study of resource-rich languages, thus ensuring the reference value of the latter. Thus, in order to avoid dysfunctions resulting from sentence truncation during training, we retake the class of learning scheme then proposed, which permits us to reap the maximum from the training process. 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:

$$\mathcal{D}_i := [\![\mathcal{C}_i]\!], \text{ with } \mathcal{C}_1 := \mathcal{K} \text{ and } \mathcal{C}_i := \mathcal{C}_{i-1} \cup \mathcal{I}_i, \ \mathcal{I}_i \subset \mathcal{C} \setminus \mathcal{C}_{i-1}, \ \|\mathcal{I}_i\| := \sigma(i), \ \forall i \geq 2 \quad (9)$$

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

Along the same lines and 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, which can be considered conservative values since smaller and larger ones are possible. Regarding the parameters used for estimating the prediction levels, the choice again goes to $\nu = 4 \times 10^{-5}$, $\varsigma = 1$ and $\lambda = 5$. This also holds true for the selection of the regression technique used for approximating the partial learning curves and for π, which falls on the *trust region method* [83] and the power law family [95], respectively.

Taking into account that real corpora are finite, we study the prediction model within their boundaries, which implies limiting the scope in measuring the layers of convergence. We then adapt the sampling window and the control levels to the size of the corpus now considered. So, if $[\![\ell]\!]$ denotes the position of the first sentence-ending beyond the ℓ-th word, the former comprises the interval $[[\![5 \times 10^3]\!], [\![7 \times 10^5]\!]]$, whilst the latter are taken from control sequences in $[[\![3 \times 10^5]\!], [\![7 \times 10^5]\!]]$. In order to confer additional stability on our measures, we apply a *k*-fold cross validation [96] to compute the samples, with $k = 10$.

6. Discussion

As mentioned, the experiments are studied from two complementary points of view, quantitative and qualitative, according to the performance metrics previously introduced.

6.1. The Sets of Runs

To illustrate the predictability of the learning curves for the Xiada corpus, we start with a collection of runs, $\mathcal{C} = \{\mathcal{E}_i\}_{i \in I}$ generated from the data compiled in Table 1. The latter includes an entry for each one of the learners previously enumerated, together with its PLevel and CLevel, as well as the values for Ac and EAc along the control sequence, from which to calculate MAPEs, DMRs and RRs. In order to improve understanding, all the levels managed are indicated by their associated word positions in the corpus, which is denoted by using a superscript **wp** in their identification labels.

One detail that attracts our attention is that the run associated with TreeTagger does not reach the PLevel within the limits of the training corpus. This behavior is certainly singular among all the taggers considered, which highlights the variety of factors that impact the evaluation of learners, and that, in this case, leads us to discard considering it in our study. In other words, in a real model selection process on the Xiada corpus

selected here, TreeTagger would not even be placed among the hypotheses that allow the application of the prediction technique considered.

Table 1. Monitoring of runs along the control sequences.

	PLevel^{wp}	τ	CLevel^{wp}	Control-Level^{wp}										MAPE	DMR	RR
				⟦3×10^5⟧		⟦4×10^5⟧		⟦5×10^5⟧		⟦6×10^5⟧		⟦7×10^5⟧				
				Ac	EAc	Ac	EAc	Ac	EAc	Ac	EAc	Ac	EAc			
fnTBL	105.003	2.40	150.017	94.16	93.87	94.57	94.30	94.96	94.61	95.16	94.84	95.34	95.03	0.32	85.71	90.00
MaxEnt	110.047	2.50	135.019	92.90	92.78	93.30	93.19	93.58	93.48	93.85	93.70	94.08	93.88	0.15	100.00	100.00
MBT	85.012	2.20	145.016	92.97	92.84	93.42	93.22	93.76	93.50	94.01	93.72	94.30	93.89	0.28	100.00	92.31
Morfette	75.011	2.60	105.003	94.61	94.54	94.98	94.89	95.21	95.14	95.41	95.33	95.55	95.49	0.09	100.00	85.71
MXPOST	110.047	2.30	145.016	93.44	93.17	93.88	93.57	94.20	93.85	94.44	94.06	94.63	94.23	0.35	100.00	100.00
Stanford	95.015	2.40	125.001	94.41	94.43	94.78	94.80	95.07	95.07	95.26	95.27	95.41	95.43	0.02	85.71	85.71
SVMTool	250.012	2.20	250.012	95.00	95.05	95.36	95.44	95.60	95.71	95.78	95.93	95.92	96.10	0.12	100.00	86.67
TnT	85.012	2.00	130.003	94.47	94.38	94.79	94.70	95.05	94.93	95.23	95.10	95.35	95.23	0.12	71.43	100.00
TreeTagger	–	2.10	–	93.36	–	93.77	–	94.02	–	94.28	–	94.42	–	–	–	–

6.2. The Quantitative Study

Our reference metric is now the MAPE, whose values are shown graphically in Figure 4 from the data compiled in Table 1. Taking into account that we are interested in numbers as small as possible, the scores range from 0.02 for Stanford to 0.35 for MXPOST in the interval $[\![⟦3×10^5⟧, ⟦7×10^5⟧]\!]$. Those results are illustrated in Figure 5, showing the learning curves and learning trends used for prediction on the runs with best and worst MAPE on the control sequence. As we have already done, the observations are generated considering the portion of the corpus taken from its beginning up to the word position indicated on the horizontal axis. Finally, 50% of MAPE values in this set of runs are in the interval $[0, 0.12]$, a proportion that reaches 75% in $[0, 0.28]$. Although these results are slightly worse than the ones reported in [70] for resource-rich languages, they are still very promising, which leads us to argue for the goodness of the proposal on the quantitative plane.
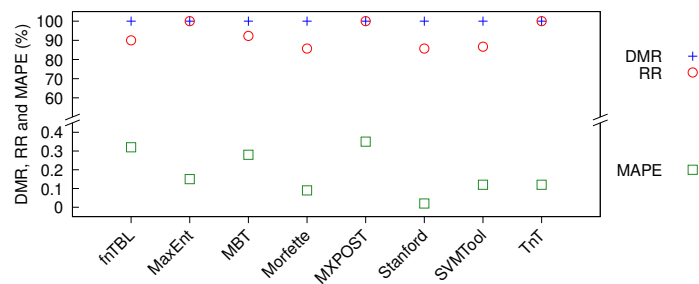

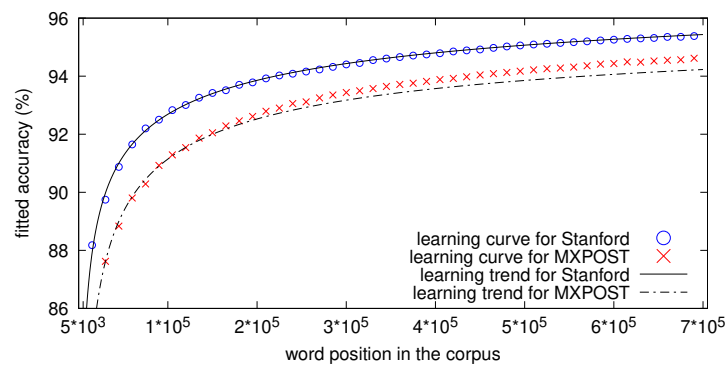**Figure 4.** MAPEs, RRs and DMRs for runs.

Figure 5. Learning trends for the best and worst MAPEs.

6.3. The Qualitative Study

Our reference metric is here the DMR, whose values are shown graphically in Figure 4 from the data compiled in Table 1. Taking into account that we are now interested in

scores close to 100, these range from 71.43 to 100, with 85.71% of the values in the interval [85.71, 100]. Moreover, the DMRs lower than 100 are the result of the intersection between the TnT learning curve with those of Stanford and fnTBL. Under these conditions, the maximum value would only be possible if the error in the estimate of the intersection points was lower than the distance between its neighboring control levels, an unrealistic proposition, given how short that distance is (5000 words). In any case, the results are comparable to those reported in [70] for resource-rich languages, also meeting our expectations from a qualitative point of view.

6.4. The Study of Robustness

The reference metric is now RR, and we are interested in values as close as possible to 100, the maximum. The results are shown in Figure 4 from the data compiled in Table 1. While RR values range from 85.71 to 100, the latter is only reached in 37.50% of the runs. This percentage rises to 62.50% for RRs in the interval [90, 100]. Overall, these results even exceed those reported in [70] for resource-rich languages, illustrating once again the good performance of the prediction model, this time against variations in its working hypotheses.

7. Conclusions and Future Work

Our proposal arises as a response to the challenge of evaluating POS tagging models in low-resource scenarios, for which non-deep learning approaches have often proven to be better suited. For this purpose, we reuse a formally correct proposal, based on the early estimation of learning curves. Technically described as the uniform convergence of a sequence of partial predictors which iteratively approximates the solution, the method acts as a proximity condition that halts the training process once a convergence/error threshold fixed by the user is reached, and has already demonstrated its validity when the availability of large enough learning datasets is not a problem. In order to ensure the reliability of the results obtained, we once again used the testing frame considered then, involving both quantitative and qualitative aspects, but also the survey of robustness against possible irregularities in the learning process.

Special attention was paid to the selection of a case study combining representativeness and access to validation resources, which is somewhat contradictory in the context under consideration. We then focus on Galician, a minority language of complex morphology, for which the collection of available training resources is reduced to a single corpus of sufficient size and quality to ensure both the validation phase and a rapid convergence process. This set of unique features allows us to simulate and evaluate short training sessions in a non-trivial learning environment, associating them with a language with important deficiencies in terms of computational resources. The results corroborate the expectations for the theoretical basis, placing the performance at a level similar to that observed in the state of the art for resource-rich languages for the same learners. This supports the effectiveness of the approach for the model selection considered and its suitability to low-resource scenarios, as initially argued.

To the best of our knowledge and belief, not only is this the first time that a proposal for estimating the performance based on the prediction of learning curves has demonstrated its feasibility in a framework of this nature, but it has done so without any type of prior specific adaptation. In other words, no operational limitations to the original conceptual design were observed. All this justifies the interest in highlighting the independence, both in terms of language and usage, of the technology deployed. A comprehensible way of doing it is extending our analysis, first to a broader set of languages in a variety of language families, and then to other fundamental and applicative NLP tasks, which establishes a clear line of future work.

Abbreviations

The following abbreviations are used in this manuscript:

Ac	Accuracy
AL	Active Learning
CLevel	Convergence Level
DL	Deep Learning
DMR	Decision-Making Reliability
EAc	Estimated Accuracy
HMM	Hidden Márkov Model
MEM	Maximum Entropy Model
MBT	Memory-Based Tagger
ML	Machine Learning
NLP	Natural Language Processing
PE	Percentage Error
POS	Part-of-Speech
PLevel	Prediction Level
RE	Reliability Estimation
RER	Reliability Estimation Ratio
RR	Robustness Rate
SVM	Support Vector Machine
WLevel	Working Level

References

1. Chiche, A.; Yitagesu, B. Part of speech tagging: A systematic review of deep learning and machine learning approaches. *J. Big Data* **2022**, *9*, 10. [CrossRef]
2. Darwish, K.; Mubarak, H.; Abdelali, A.; Eldesouki, M. Arabic POS Tagging: Don't Abandon Feature Engineering Just Yet. In Proceedings of the Third Arabic Natural Language Processing Workshop, Valencia, Spain, 3 April 2017; Association for Computational Linguistics: Madison, WI, USA, 2017; pp. 130–137.
3. Pylypenko, D.; Amponsah-Kaakyire, K.; Dutta Chowdhury, K.; van Genabith, J.; España-Bonet, C. Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Madison, WI, USA, 2021; pp. 8596–8611.
4. Tayyar Madabushi, H.; Lee, M. High Accuracy Rule-based Question Classification using Question Syntax and Semantics. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 13–16 December 2016; Association for Computational Linguistics: Madison, WI, USA, 2016; pp. 1220–1230.
5. Zhang, B.; Su, J.; Xiong, D.; Lu, Y.; Duan, H.; Yao, J. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Madison, WI, USA, 2015; pp. 2230–2235.
6. Chiong, R.; Wei, W. Named Entity Recognition Using Hybrid Machine Learning Approach. In Proceedings of the 5th IEEE International Conference on Cognitive Informatics, Beijing, China, 17–19 July 2006; IEEE CS Press: Washington, DC, USA, 2006; Volume 1, pp. 578–583.
7. Kim, J.; Ko, Y.; Seo, J. A Bootstrapping Approach With CRF and Deep Learning Models for Improving the Biomedical Named Entity Recognition in Multi-Domains. *IEEE Access* **2019**, *7*, 70308–70318. [CrossRef]

8.   Li, J.; Li, R.; Hovy, E. Recursive Deep Models for Discourse Parsing. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Madison, WI, USA, 2014; pp. 2061–2069.

9.   Crammer, K. Advanced Online Learning for Natural Language Processing. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Tutorial Abstracts, Columbus, OH, USA, 15–20 June 2008; Association for Computational Linguistics: Madison, WI, USA, 2008; p. 4.

10. Vlachos, A. Evaluating unsupervised learning for natural language processing tasks. In Proceedings of the First workshop on Unsupervised Learning in NLP, Edinburgh, UK, 30 July 2011; Association for Computational Linguistics: Madison, WI, USA, 2011; pp. 35–42.

11. Florian, R.; Hassan, H.; Jing, H.; Kambhatla, N.; Luo, X.; Nicolov, N.; Roukos, S. A Statistical Model for Multilingual Entity Detection and Tracking. In Proceedings of the Human Language Technologies Conference 2004, Boston, MA, USA, 2–7 May 2004; Association for Computational Linguistics: Madison, WI, USA, 2004; pp. 1–8.

12. Xue, G.R.; Dai, W.; Yang, Q.; Yu, Y. Topic-bridged PLSA for Cross-domain Text Classification. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 20–24 July 2008; ACM Press: New York, NY, USA, 2008; pp. 627–634.

13. Chan, S.; Honari Jahromi, M.; Benetti, B.; Lakhani, A.; Fyshe, A. Ensemble Methods for Native Language Identification. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark, 8 September 2017; Association for Computational Linguistics: Madison, WI, USA, 2017; pp. 217–223.

14. Libovický, J.; Helcl, J. End-to-End Non-Autoregressive Neural Machine Translation with Connectionist Temporal Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2–4 November 2018; Association for Computational Linguistics: Madison, WI, USA, 2018; pp. 3016–3021.

15. Cortes, E.; Woloszyn, V.; Binder, A.; Himmelsbach, T.; Barone, D.; Möller, S. An Empirical Comparison of Question Classification Methods for Question Answering Systems. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020; European Language Resources Association: Paris, France, 2020; pp. 5408–5416.

16. Swier, R.S.; Stevenson, S. Unsupervised Semantic Role Labeling. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Madison, WI, USA, 2004; pp. 95–102.

17. Glorot, X.; Bordes, A.; Bengio, Y. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; ACM Press: New York, NY, USA, 2011; pp. 513–520.

18. Dai, W.; Xue, G.R.; Yang, Q.; Yu, Y. Transferring Naive Bayes Classifiers for Text Classification. In Proceedings of the 22nd National Conference on Artificial Intelligence, Vancouver, BC, Canada, 22–26 July 2007; AAAI Press: Menlo Park, CA, USA, 2007; Volume 1, pp. 540–545.

19. Ebrahimi, M.; Eberhart, A.; Bianchi, F.; Hitzler, P. Towards Bridging the Neuro-Symbolic Gap: Deep Deductive Reasoners. *Appl. Intell.* **2021**, *51*, 6326–6348. [CrossRef]

20. Poggio, T.; Banburski, A.; Liao, Q. Theoretical issues in deep networks. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 30039–30045. [CrossRef] [PubMed]

21. Hao, H.; Mengya, G.; Mingsheng, W. Relieving the Incompatibility of Network Representation and Classification for Long-Tailed Data Distribution. *Comput. Intell. Neurosci.* **2021**, *2021*, 6702625,

22. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep Long-Tailed Learning: A Survey. *arXiv* **2021**, arXiv:2110.04596.

23. Hoefler, T.; Alistarh, D.T.B.N.N.D.; Peste, A. Analytically Tractable Hidden-States Inference in Bayesian Neural Networks. *J. Mach. Learn. Res.* **2021**, *23*, 1–124.

24. Li, H. Deep learning for natural language processing: Advantages and challenges. *Natl. Sci. Rev.* **2017**, *5*, 24–26. [CrossRef]

25. Hedderich, M.A.; Lange, L.; Adel, H.; Strötgen, J.; Klakow, D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; Association for Computational Linguistics: Madison, WI, USA, 2021; pp. 2545–2568.

26. Chakrabarty, A.; Chaturvedi, A.; Garain, U. NeuMorph: Neural Morphological Tagging for Low-Resource Languages— An Experimental Study for Indic Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2019**, *19*, 1–19. doi: 10.1145/3342354. [CrossRef]

27. Geman, S.; Bienenstock, E.; Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Comput.* **1992**, *4*, 1–58. [CrossRef]

28. Magnini, B.; Lavelli, A.; Magnolini, S. Comparing Machine Learning and Deep Learning Approaches on NLP Tasks for the Italian Language. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Paris, France, 2020; pp. 2110–2119.

29. Saied, H.A.; Candito, M.; Constant, M. Comparing linear and neural models for competitive MWE identification. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland, 30 September–2 October 2019; Linköping University Electronic Press: Linköping, Sweden, 2019; pp. 86–96.

30. Wang, M.; Manning, C.D. Effect of Non-linear Deep Architecture in Sequence Labeling. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–18 October 2013; pp. 1285–1291.

31. Song, H.J.; Son, J.W.; Noh, T.G.; Park, S.B.; Lee, S.J. 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, Jeju Island, Korea, 8–14 July 2012; Association for Computational Linguistics: Madison, WI, USA, 2012; Volume 1, pp. 1025–1034.

32. Hoesen, D.; Purwarianti, A. Investigating Bi-LSTM and CRF with POS Tag Embedding for Indonesian Named Entity Tagger. *ar*Xiv **2020**, *ar*Xiv:2009.05687.

33. Khan, W.; Daud, A.; Khan, K.; Nasir, J.A.; Basheri, M.; Aljohani, N.; Alotaibi, F.S. Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches. *IEEE Access* **2019**, *7*, 38918–38936. [CrossRef]

34. Ljubešić, N. Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, Santa Fe, NM, USA, 20–21 August 2018; Association for Computational Linguistics: Madison, WI, USA, 2018; pp. 156–163.

35. Stankovic, R.; Šandrih, B.; Krstev, C.; Utvić, M.; Skoric, M. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; European Language Resources Association: Paris, France, 2022; pp. 3954–3962.

36. Todi, K.K.; Mishra, P.; Sharma, D.M. Building a Kannada POS Tagger Using Machine Learning and Neural Network Models. *arXiv* **2018**, arXiv:1808.03175.

37. Murata, N.; Yoshizawa, S.; ichi Amari, S. Learning Curves, Model Selection and Complexity of Neural Networks. In *Neural Information Processing Systems*; Hanson, S.J., Cowan, D., Giles, C.L., Eds.; Morgan Kaufmann: San Mateo, CA, USA, 1993; Volume 5, pp. 607–614.

38. Bertoldi, N.; Cettolo, M.; Federico, M.; Buck, C. Evaluating the Learning Curve of Domain Adaptive Statistical Machine Translation Systems. In Proceedings of the 7th Workshop on Statistical Machine Translation, Montreal, Canada, 7–8 June 2012; Association for Computational Linguistics: Madison, WI, USA, 2012; pp. 433–441.

39. Turchi, M.; De Bie, T.; Cristianini, N. Learning Performance of a Machine Translation System: A Statistical and Computational Analysis. In Proceedings of the 3rd Workshop on Statistical Machine Translation, Columbus, OH, USA, 19 June 2008; Association for Computational Linguistics: Madison, WI, USA, 2008; pp. 35–43.

40. Koehn, P.; Och, F.J.; Marcu, D. 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, Edmonton, AB, Canada, 28–30 May 2003; Association for Computational Linguistics: Madison, WI, USA, 2003; Volume 1, pp. 48–54.

41. Kolachina, P.; Cancedda, N.; Dymetman, M.; Venkatapathy, S. Prediction of Learning Curves in Machine Translation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, Jeju Island, Korea, 8–14 July 2012; Association for Computational Linguistics: Madison, WI, USA, 2012; Volume 1, pp. 22–30.

42. Birch, A.; Osborne, M.; Koehn, P. Predicting Success in Machine Translation. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Madison, WI, USA, 2008; pp. 745–754.

43. Cohn, D.; Atlas, L.; Ladner, R. Improving Generalization with Active Learning. *Mach. Learn.* **1994**, *15*, 201–221. [CrossRef]

44. Culotta, A.; McCallum, A. Reducing Labeling Effort for Structured Prediction Tasks. In Proceedings of the 20th National Conference on Artificial Intelligence, Pittsburgh, PA, USA, 9–13 July 2005; AAAI Press: Essex, UK, 2005; Volume 2, pp. 746–751.

45. Thompson, C.A.; Califf, M.E.; Mooney, R.J. Active Learning for Natural Language Parsing and Information Extraction. In Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 27–30 June 1999; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; pp. 406–414.

46. Becker, M.; Osborne, M. A Two-stage Method for Active Learning of Statistical Grammars. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, UK, 30 July–5 August 2005; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2005; pp. 991–996.

47. Tang, M.; Luo, X.; Roukos, S. Active Learning for Statistical Natural Language Parsing. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Madison, WI, USA, 2002; pp. 120–127.

48. Lewis, D.D.; Gale, W.A. 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, Ireland, 3–6 July 1994; Springer: Berlin/Heidelberg, Germany, 1994; pp. 3–12.

49. Liere, R.; Tadepalli, P. Active learning with committees for text categorization. In Proceedings of the 14th National Conference on Artificial Intelligence, Providence, RI, USA, 27–31 July 1997; AAAI Press: Essex, UK, 1997; pp. 591–596.

50. McCallum, A.; Nigam, K. Employing EM and Pool-Based Active Learning for Text Classification. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, USA, 24–27 July 1998; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1998; pp. 350–358.

51. Tong, S.; Koller, D. Support Vector Machine Active Learning with Applications to Text Classification. *J. Mach. Learn. Res.* **2002**, *2*, 45–66.

52. Dagan, I.; Engelson, S.P. Committee-Based Sampling For Training Probabilistic Classifiers. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp. 150–157.

53. Haertel, R.; Ringger, E.; Seppi, K.; Carroll, J.; McClanahan, P. 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, OH, USA, 15–20 June 2008; Association for Computational Linguistics: Madison, WI, USA, 2008; pp. 65–68.

54. Ringger, E.; McClanahan, P.; Haertel, R.; Busby, G.; Carmen, M.; Carroll, J.; Seppi, K.; Lonsdale, D. Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation. In Proceedings of the Linguistic Annotation Workshop, Prague, Czech Republic, 28–29 June 2007; Association for Computational Linguistics: Madison, WI, USA, 2007; pp. 101–108.

55. Laws, F.; Schütze, H. Stopping Criteria for Active Learning of Named Entity Recognition. In Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, 18–22 August 2008; Association for Computational Linguistics: Madison, WI, USA, 2008; Volume 1, pp. 465–472.

56. Shen, D.; Zhang, J.; Su, J.; Zhou, G.; Tan, C.L. Multi-criteria-based Active Learning for Named Entity Recognition. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004; Association for Computational Linguistics: Madison, WI, USA, 2004; pp. 589–596.

57. Tomanek, K.; Wermter, J.; Hahn, U. 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, Czech Republic, 28–30 June 2007; Association for Computational Linguistics: Madison, WI, USA, 2007; pp. 486–495.

58. Chan, Y.S.; Ng, H.T. Domain Adaptation with Active Learning for Word Sense Disambiguation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 28–30 June 2007; Association for Computational Linguistics: Madison, WI, USA, 2007; pp. 49–56.

59. Chen, J.; Schein, A.; Ungar, L.; Palmer, M. 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, NY, USA, 4–9 June 2006; Association for Computational Linguistics: Madison, WI, USA, 2006; pp. 120–127.

60. Zhu, J.; Hovy, E. 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, Czech Republic, 28–30 June 2007; Association for Computational Linguistics: Madison, WI, USA, 2007; pp. 783–790.

61. Baldridge, J.; Osborne, M. Active learning and logarithmic opinion pools for HPSG parse selection. *Nat. Lang. Eng.* **2008**, *14*, 191–222. [CrossRef]

62. Ein-Dor, L.; Halfon, A.; Gera, A.; Shnarch, E.; Dankin, L.; Choshen, L.; Danilevsky, M.; Aharonov, R.; Katz, Y.; Slonim, N. Active Learning for BERT: An Empirical Study. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; Association for Computational Linguistics: Online, 2020; pp. 7949–7962.

63. Liu, M.; Buntine, W.; Haffari, G. Learning to Actively Learn Neural Machine Translation. In Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, 31 October–1 November 2018; Association for Computational Linguistics: Madison, WI, USA, 2018; pp. 334–344.

64. Lowell, D.; Lipton, Z.C.; Wallace, B.C. Practical Obstacles to Deploying Active Learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Madison, WI, USA, 2019; pp. 21–30.

65. Anastasopoulos, A.; Lekakou, M.; Quer, J.; Zimianiti, E.; DeBenedetto, J.; Chiang, D. Part-of-Speech Tagging on an Endangered Language: A Parallel Griko-Italian Resource. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; Association for Computational Linguistics: Madison, WI, USA, 2018; pp. 2529–2539.

66. Chaudhary, A.; Anastasopoulos, A.; Sheikh, Z.; Neubig, G. Reducing Confusion in Active Learning for Part-of-Speech Tagging. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1–16.

67. Erdmann, A.; Wrisley, D.J.; Allen, B.; Brown, C.; Cohen-Bodénès, S.; Elsner, M.; Feng, Y.; Joseph, B.; Joyeux-Prunel, B.; de Marneffe, M.C. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long and Short Papers, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Madison, WI, USA, 2019; Volume 1, pp. 2223–2234.

68. Kim, Y. Deep Active Learning for Sequence Labeling Based on Diversity and Uncertainty in Gradient. In Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, Suzhou, China, 7 December 2020; Association for Computational Linguistics: Madison, WI, USA, 2020; pp. 1–8.

69. Settles, B.; Craven, M. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Madison, WI, USA, 2008; pp. 1070–1079.

70. Vilares, M.; Darriba, V.M.; Ribadas, F.J. Modeling of learning curves with applications to POS tagging. *Comput. Speech Lang.* **2017**, *41*, 1–28. [CrossRef]

71. Baker, B.; Gupta, O.; Raskar, R.; Naik, N. Accelerating neural architecture search using performance prediction. In Proceedings of the 6th International Conference on Learning Representations, ICLR'18, Vancouver, BC, Canada, 30 April–3 May 2018.

72. Domhan, T.; Springenberg, J.T.; Hutter, F. Speeding up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves. In Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentine, 25–31 July 2015; AAAI Press: Essex, UK, 2015; pp. 3460–3468.

73. Klein, A.; Falkner, S.; Springenberg, J.T.; Hutter, F. Learning Curve Prediction with Bayesian Neural Networks. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.

74. Vilares, M.; Darriba, V.; Vilares, J. Absolute convergence and error thresholds in non-active adaptive sampling. *J. Comput. Syst. Sci.* **2022**, *129*, 39–61. [CrossRef]

75. Vilares, M.; Darriba, V.M.; Vilares, J. Adaptive scheduling for adaptive sampling in POS taggers construction. *Comput. Speech Lang.* **2020**, *60*, 101020.

76. Domingo, C.; Gavaldà, R.; Watanabe, O. Adaptive Sampling Methods for Scaling Up Knowledge Discovery Algorithms. *Data Min. Knowl. Discov.* **2002**, *6*, 131–152. [CrossRef]

77. Meek, C.; Thiesson, B.; Heckerman, D. The Learning-curve Sampling Method Applied to Model-based Clustering. *J. Mach. Learn. Res.* **2002**, *2*, 397–418.

78. Mohr, F.; van Rijn, J.N. Fast and Informative Model Selection using Learning Curve Cross-Validation. *arXiv* **2021**, arXiv:2111.13914.

79. Schütze, H.; Velipasaoglu, E.; Pedersen, J.O. Performance Thresholding in Practical Text Classification. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, VA, USA, 6–11 November 2006; ACM Press: New York, NY, USA; pp. 662–671.

80. Tomanek, K.; Hahn, U. Approximating Learning Curves for Active-Learning-Driven Annotation. In Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 28–30 May 2008; European Language Resources Association: Paris, France, 2008; pp. 1319–1324.

81. Giménez, J.; Márquez, L. SVMTool: A general POS tagger generator based on support vector machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004; European Language Resources Association: Paris, France, 2004; pp. 43–46.

82. Etiquetador/Lematizador do Galego Actual (XIADA) [v2.8]—Corpus de adestramento. Centro Ramón Piñeiro para a Investigación en Humanidades. Available online: http://corpus.cirp.gal/xiada/descargas/texto_corpus (accesed on 13 July 2022).

83. Branch, M.A.; Coleman, T.F.; Li, Y. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM J. Sci. Comput.* **1999**, *21*, 1–23. [CrossRef]

84. Vandome, P. Econometric forecasting for the United Kingdom. *Bull. Oxf. Univ. Inst. Econ. Stat.* **1963**, *25*, 239–281. [CrossRef]

85. Vilares, M.; Graña, J.; Araujo, T.; Cabrero, D.; Diz, I. A tagger environment for Galician. In Proceedings of the Workshop on Language Resources for European Minority Languages, Granada, Spain, 27 May 1998; European Language Resources Association: Paris, France, 1998.

86. Brants, T. TnT: A Statistical Part-of-speech Tagger. In Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, Washington, USA, 29 April–4 May 2000; Association for Computational Linguistics: Madison, WI, USA, 2000; pp. 224–231.

87. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, 1994; Association for Computational Linguistics: Madison, WI, USA, 1994; pp. 44–49.

88. Chrupala, G.; Dinu, G.; van Genabith, J. Learning Morphology with Morfette. In Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 28–30 May 2008; European Language Resources Association: Paris, France, 2008; pp. 2362–2367.

89. Collins, M. 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, Philadelphia, PA, USA, 6–7 July 2002; Association for Computational Linguistics: Madison, WI, USA, 2002; Volume 10, pp. 1–8.

90. Ratnaparkhi, A. A Maximum Entropy Model for Part-Of-Speech Tagging. In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 17–18 May 1996; Association for Computational Linguistics: Madison, WI, USA, 1996; pp. 133–142.

91. Toutanova, K.; Klein, D.; Manning, C.D.; Singer, Y. 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, Edmonton, AB, Canada, 28–30 May 2003; Association for Computational Linguistics: Madison, WI, USA, 2003; Volume 1, pp. 173–180.

92. Ngai, G.; Florian, R. 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, PA, USA, 2–7 June 2001; Association for Computational Linguistics: Madison, WI, USA, 2001; pp. 1–8.

93. Brill, E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Comput. Linguist.* **1995**, *21*, 543–565.

94. Daelemans, W.; Zavrel, J.; Berck, P.; Gillis, S. MBT: A Memory–Based Part-of-speech Tagger Generator. In Proceedings of the 4th Workshop on Very Large Corpora, Herstmonceux Castle, Sussex, UK, 4 August 1996; Association for Computational Linguistics: Madison, WI, USA, 1996; pp. 14–27.

95. Gu, B.; Hu, F.; Liu, H. Modelling Classification Performance for Large Data Sets. In Proceedings of the 2nd International Conference on Advances in Web-Age Information Management, Xi'an, China, 9–11 July 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 317–328.

96. Clark, A.; Fox, C.; Lappin, S. *The Handbook of Computational Linguistics and Natural Language Processing*; John Wiley & Sons: Hoboken, NJ, USA, 2010.