Diversity-Aware Batch Active Learning for Dependency Parsing

Tianze Shi[*]
Cornell University
tianze@cs.cornell.edu

Adrian Benton
Bloomberg L.P.
abenton10@bloomberg.net

Igor Malioutov
Bloomberg L.P.
imalioutov@bloomberg.net

Ozan İrsoy
Bloomberg L.P.
oirsoy@bloomberg.net

Abstract

While the predictive performance of modern statistical dependency parsers relies heavily on the availability of expensive expert-annotated treebank data, not all annotations contribute equally to the training of the parsers. In this paper, we attempt to reduce the number of labeled examples needed to train a strong dependency parser using batch active learning (AL). In particular, we investigate whether enforcing diversity in the sampled batches, using determinantal point processes (DPPs), can improve over their diversity-agnostic counterparts. Simulation experiments on an English newswire corpus show that selecting diverse batches with DPPs is superior to strong selection strategies that do not enforce batch diversity, especially during the initial stages of the learning process. Additionally, our diversity-aware strategy is robust under a corpus duplication setting, where diversity-agnostic sampling strategies exhibit significant degradation.

1 Introduction

Though critical to parser training, data annotations for dependency parsing are both expensive and time-consuming to obtain. Syntactic analysis requires linguistic expertise and even after extensive training, data annotation can still be burdensome. The Penn Treebank project (Marcus et al., 1993) reports that after two months of training, the annotators average 750 tokens per hour on the bracketing task; the Prague Dependency Treebank (Böhmová et al., 2003) cost over \$600,000 and required 5 years to annotate roughly 90,000 sentences (over \$5 per sentence). These high annotation costs present a significant challenge to developing accurate dependency parsers for under-resourced languages and domains.

Active learning (AL; Settles, 2009) is a promising technique to reduce the annotation effort required to train a strong dependency parser by intelligently selecting samples to annotate such that the return of each annotator hour is as high as possible. Popular selection strategies, such as uncertainty sampling, associate each instance with a *quality* measure based on the uncertainty or confidence level of the current parser, and higher-quality instances are selected for annotation.

We focus on *batch mode* AL, since it is generally more efficient for annotators to label in bulk. While early work in AL for parsing (Tang et al., 2002; Hwa, 2000, 2004) cautions against using individually-computed quality measures in the batch setting, more recent work demonstrates empirical success (e.g., Li et al., 2016) without explicitly handling intra-batch *diversity*. In this paper, we explore whether a diversity-aware approach can improve the state of the art in AL for dependency parsing. Specifically, we consider samples drawn from determinantal point processes (DPPs) as a query strategy to select batches of high-quality, yet dissimilar instances (Kulesza and Taskar, 2012).

In this paper, we (1) propose a diversity-aware batch AL query strategy for dependency parsing compatible with existing selection strategies, (2) empirically study three AL strategies with and without diversity factors, and (3) find that diversity-aware selection strategies are superior to their diversity-agnostic counterparts, especially during the early stages of the learning process, in simulation experiments on an English newswire corpus. This is critical in low-budget AL settings, which we further confirm in a corpus duplication setting.[1]

2 Active Learning for Dependency Parsing

2.1 Dependency Parsing

Dependency parsing (Kübler et al., 2008) aims to find the syntactic dependency structure, y, given a length-n input sentence $x = x_1, x_2, \ldots, x_n$, where

[*] Work done during an internship at Bloomberg L.P.

[1] Our code is publicly available at https://github.com/tzshi/dpp-al-parsing-naacl21.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2616–2626
June 6–11, 2021. ©2021 Association for Computational Linguistics

y is a set of n arcs over the tokens and the dummy root symbol x_0, and each arc $(h, m) \in y$ specifies the head, h, and modifier word, m.[2] In this work, we adopt the conceptually-simple edge-factored deep biaffine dependency parser (Dozat and Manning, 2017), which is competitive with the state of the art in terms of accuracy, The parser assigns a locally-normalized attachment probability $P_{\text{att}}(\text{head}(m) = h \mid x)$ to each attachment candidate pair (h, m) based on a biaffine scoring function. Refer to Appendix A for architecture details.

We define the score of the candidate parse tree $s(y \mid x)$ as $\sum_{(h,m)\in y} \log P_{\text{att}}(\text{head}(m) = h \mid x)$. The decoder finds the best scoring $\hat{y}$ among all valid trees $\mathcal{Y}(x)$: $\hat{y} = \arg\max_{y\in\mathcal{Y}(x)} s(y \mid x)$.

2.2 Active Learning (AL)

We consider the pool-based batch AL scenario where we assume a large collection of unlabeled instances $\mathcal{U}$ from which we sample a small subset at a time to annotate after each round to form an expanding labeled training set $\mathcal{L}$ (Lewis and Gale, 1994). We use the superscript i to denote the pool of instances $\mathcal{U}^i$ and $\mathcal{L}^i$ after the i-th round. $\mathcal{L}^0$ is a small set of seed labeled instances to initiate the process. Each iteration starts with training a model $\mathcal{M}^i$ based on $\mathcal{L}^i$. Next, all unlabeled data instances in $\mathcal{U}^i$ are parsed by $\mathcal{M}^i$ and we select a batch $\mathcal{U}'$ to annotate based on some criterion $\mathcal{U}' = \mathcal{C}(\mathcal{M}^i, \mathcal{U}^i)$. The resulting labeled subset $\mathcal{L}'$ is added to $\mathcal{L}^{i+1} = \mathcal{L}^i \bigcup \mathcal{L}'$ and $\mathcal{U}^{i+1} = \mathcal{U}^i - \mathcal{U}'$.

The definition of the selection criterion $\mathcal{C}$ is critical. A typical strategy associates each unlabeled instance $\mathcal{U}_i$ with a quality measure q_i based on, for example, the model uncertainty level when parsing $\mathcal{U}_i$. A *diversity-agnostic* criterion sorts all unlabeled instances by their quality measures and takes the top-k as $\mathcal{U}'$ for a budget k.

2.3 Quality Measures

We consider three commonly-used quality measures adapted to the task of dependency parsing, including uncertainty sampling, Bayesian active learning, and a representativeness-based strategy.

Average Marginal Probability (AMP) measures parser uncertainty (Li et al., 2016):

$$\mathsf{AMP} = 1 - \frac{1}{n} \sum_{(\hat{h},m)\in\hat{y}} P_{\text{mar}}(\text{head}(m) = \hat{h} \mid x),$$

where P_{mar} is the marginal attachment probability

$$P_{\text{mar}}(\text{head}(m) = h \mid x) = \sum_{(h,m)\in y} P(y \mid x),$$

and $P(y \mid x) = \frac{\exp(s(y|x))}{\sum_{y'\in\mathcal{Y}(x)} \exp(s(y'|x))}$. The marginal probabilities can be derived efficiently using Kirchhoff's theorem (Tutte, 1984; Koo et al., 2007).

Bayesian Active Learning by Disagreement (BALD) measures the mutual information between the model parameters and the predictions. We adopt the Monte Carlo dropout-based variant (Gal et al., 2017; Siddhant and Lipton, 2018) and measure the disagreement among predictions from a neural model with K different dropout masks, which has been applied to active learning in NLP. We adapt BALD to dependency parsing by aggregating disagreement at a token level:

$$\mathsf{BALD} = 1 - \frac{1}{n} \sum_{m} \frac{\text{count}(\text{mode}(h_m^1, \ldots, h_m^K))}{K},$$

where h_m^k denotes that (h_m^k, m) appears in the prediction given by the k-th model.

Information Density (ID) mitigates the tendency of uncertainty sampling to favor outliers by weighing examples by how *representative* they are of the entire dataset (Settles and Craven, 2008):

$$\mathsf{ID} = \mathsf{AMP} \times \left(\frac{1}{|\mathcal{U}|} \sum_{x'\in\mathcal{U}} \text{sim}_{\cos}(x, x') \right),$$

where cosine similarity is computed from the averaged contextualized features (§3.2).

2.4 Learning from Partial Annotations

We follow Li et al. (2016) and select tokens to annotate their heads instead of annotating full sentences. We first pick the most informative sentences and then choose $p\%$ tokens from them based on the token-level versions of the quality measures (e.g., marginal probability instead of AMP).

3 Selecting Diverse Samples

Near-duplicate examples are common in real-world data (Broder et al., 1997; Manku et al., 2007), but they provide overlapping utility to model training. In the extreme case, with a diversity-agnostic strategy for active learning, identical examples will be selected/excluded at the same time (Hwa, 2004). To address this issue and to best utilize the annotation budget, it is important to consider diversity. We adapt Bıyık et al. (2019) to explicitly model diversity using determinantal point processes (DPPs).

3.1 Determinantal Point Processes

A DPP defines a probability distribution over subsets of some ground set of elements (Kulesza, 2012). In AL, the ground set is the unlabeled

[2] For clarity, here we describe unlabeled parsing. In our experiments, we train labeled dependency parsers, which additionally predict a dependency relation label l for each arc.

pool $\mathcal{U}$ and a subset corresponds to a batch of instances $\mathcal{U}'$ drawn from $\mathcal{U}$. DPPs provide an explicit mechanism to ensure high-quality yet diverse sample selection by modeling both the quality measures and the similarities among examples. We adopt the L-ensemble representation of DPPs using the quality-diversity decomposition (Kulesza and Taskar, 2012) and parameterize the matrix L as $L_{ij} = q_i \phi_i \phi_j^T q_j$, where each $q_i \in \mathbb{R}$ is the quality measure for $\mathcal{U}_i$ and each $\phi_i \in \mathbb{R}^{1 \times d}$ is a d-dimensional vector representation of $\mathcal{U}_i$, which we refer to as $\mathcal{U}_i$'s *diversity features*.[3] The probability of selecting a batch B is given by $P(B \subseteq \mathcal{U}) \propto \det(L_B)$, where $\det(\cdot)$ calculates the determinant and L_B is the submatrix of L indexed by elements in B.

DPPs place high probability on diverse subsets of high-quality items. Intuitively, the determinant of L_B corresponds to the volume spanned by the set of vectors $\{q_i \phi_i \mid i \in B\}$, and subsets with larger q values and orthogonal ϕ vectors span larger volumes than those with smaller q values or similar ϕ vectors. We follow Kulesza (2012) and adapt their greedy algorithm for finding the approximate mode $\arg\max_B P(B \subseteq \mathcal{U})$. This algorithm is reproduced in Algorithm E1 in the appendix.

3.2 Diversity Features

We consider two possibilities for the diversity features ϕ. Each feature vector is unit-normalized.

Averaged Contextualized Features are defined as $\frac{1}{n} \sum_i \mathbf{x}_i$, where $\mathbf{x}_i$ is a contextualized vector of x_i from the feature extractor used by the parser. By this definition, we consider the instances to be similar to each other when the neural feature extractor returns similar features such that the parser is likely to predict similar structures for these instances.

Predicted Subgraph Counts explicitly represent the predicted tree structure. To balance richness and sparsity, we count the labeled but unlexicalized subgraph formed by the grandparent, the parent and the token itself. Specifically, for each token m, we can extract a subgraph denoted by (r_1, r_2), assuming the predicted dependency relation between its grandparent g and its parent h is r_1, and the relation between h and m is r_2. The parse tree for a length-n sentence contains n such subgraphs. We apply tf-idf weighting to discount the influence from frequent subgraphs.

Batch	5		10	
Strategy	w/o DPP	w/ DPP	w/o DPP	w/ DPP
Random	$85.68_{\pm.26}$	$\mathbf{86.61}_{\pm.28}$	$87.84_{\pm.26}$	$\mathbf{88.55}_{\pm.23}$
AMP	$85.98_{\pm.22}$	$\mathbf{86.77}_{\pm.43}$	$88.80_{\pm.18}$	$\mathbf{89.23}_{\pm.29}$
BALD	$86.24_{\pm.40}$	$\mathbf{86.86}_{\pm.31}$	$88.66_{\pm.36}$	$\mathbf{89.03}_{\pm.10}$
ID	$\mathbf{86.68}_{\pm.26}$	$86.56_{\pm.24}$	$88.96_{\pm.20}$	$\mathbf{89.06}_{\pm.16}$

Table 1: LAS after 5 and 10 rounds of annotation for strategies with and without modeling diversity through DPP.

4 Experiments and Results

Dataset We use the Revised English News Text Treebank[4] (Bies et al., 2015) converted to Universal Dependencies 2.0 using the conversion tool included in Stanford Parser (Manning et al., 2014) version 4.0.0. We use sections 02-21 for training, 22 for development and 23 for test.

Setting We perform experiments by simulating the annotation process using treebank data. We sample 128 sentences uniformly for the initial labeled pool and each following round selects 500 tokens for partial annotation. We run each setting five times using different random initializations and report the means and standard deviations of the labeled attachment scores (LAS). Appendix B has unlabeled attachment score (UAS) results.

Baselines While we construct our own baselines for self-contained comparisons, the diversity-agnostic AMP (w/o DPP) largely replicates the state-of-the-art selection strategy of Li et al. (2016).

Implementation We finetune a pretrained multilingual XLM-RoBERTa base model (Conneau et al., 2020) as our feature extractor.[5] See Appendix E for implementation details.

Main Results Table 1 compares LAS after 5 and 10 rounds of annotation. Our dependency parser reaches 95.64 UAS and 94.06 LAS, when trained with the full dataset (more than one million tokens). Training data collected from 30 annotation rounds ($\approx$ 17,500 tokens) correspond to roughly 2% of the full dataset, but already support an LAS of up to 92 through AL. We find that diversity-aware strategies generally improve over their diversity-agnostic counterparts. Even for a random selection strategy, ensuring diversity with a DPP is superior

[3] Although certain applications of DPPs may learn q and ϕ representations from supervision, we define q and ϕ *a priori*, since acquiring supervision in AL is, by definition, expensive.

[4] https://catalog.ldc.upenn.edu/LDC2015T13

[5] To construct the averaged contextualized features, we also use the fine-tuned feature extractor. In our preliminary experiments, we have tried freezing the feature extractors, but this variant did not perform as well.

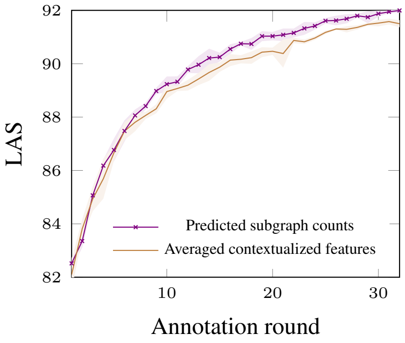

Figure 1: Learning curves for our DPP-based diversity-aware selection strategies, comparing predicted subgraph counts versus averaged contextualized features as diversity features. Both use AMP as their quality measures.

to simple random selection. With AMP and BALD, our diversity-aware strategy sees a larger improvement earlier in the learning process. ID models representativeness of instances, and our diversity-aware strategy adds less utility compared with other quality measures, although we do notice a large improvement after the first annotation round for ID: $82.40_{\pm.48}$ vs. $83.36_{\pm.54}$ (w/ DPP) – a similar trend to AMP and BALD, but at an earlier stage of AL.

Experiments with Different Diversity Features Figure 1 compares our two definitions of diversity features, and we find that predicted subgraph counts provide stronger performance than that of averaged contextualized features. We hypothesize this is due to the fact that the subgraph counts represent structures more explicitly, thus they are more useful in maintaining structural diversity in AL.

Intra-Batch Diversity To quantify intra-batch diversity among the set of sentences B picked by the selection strategies, we adapt the measures used by Chen et al. (2018) and define intra-batch average distance (IBAD) and intra-batch minimal distance (IBMD) as follows:

$$\text{IBAD} = \underset{i,j\in B, i\neq j}{\text{mean}} (1 - \text{sim}_{\cos}(i,j)),$$
$$\text{IBMD} = \underset{i\in B}{\text{mean}} \min_{j\in B, i\neq j} (1 - \text{sim}_{\cos}(i,j)).$$

A higher value on these measures indicates better intra-batch diversity. Figure 2 compares diversity-agnostic and diversity-aware sampling strategies using the two different diversity features. We confirm that DPPs indeed promote diverse samples in the selected batches, while intra-batch diversity naturally increases even for the diversity-agnostic strategies. Additionally, we observe that the benefits of DPPs are more prominent when using predicted subgraph counts compared with averaged contextualized features. This can help explain the relative success of the former diversity features.

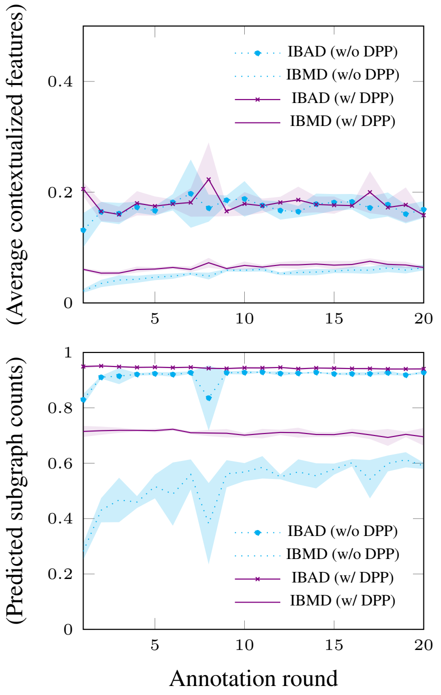

Figure 2: Intra-batch average distance (IBAD) and intra-batch minimal distance (IBMD) measures comparing diversity-agnostic and diversity-aware AMP-based sample selection strategies. The distances are derived from averaged contextualized features (top) and predicted subgraph counts (bottom). A higher value indicates better intra-batch diversity.

Corpus Duplication Setting In our qualitative analysis (Appendix C), we find that diversity-agnostic selection strategies tend to select near-duplicate sentences. To examine this phenomenon in isolation, we repeat the training corpus twice and observe the effect of diversity-aware strategies. The corpus duplication technique has been previously used to probe semantic models (Schofield et al., 2017). Figure 3 shows learning curves for strategies under the original and corpus duplication settings. As expected, diversity-aware strategies consistently outperform their diversity-agnostic counterparts across both settings, while some diversity-agnostic strategies (e.g., AMP) even underperform uniform random selection in the duplicated setting.

Interpreting the Effectiveness of Diversity-Agnostic Models Figure 4 visualizes the density distributions of the top 200 data instances by AMP over the diversity feature space reduced to two dimensions through t-SNE (van der Maaten and Hinton, 2008). During the initial stage of active learning, data with the highest quality measures are con-

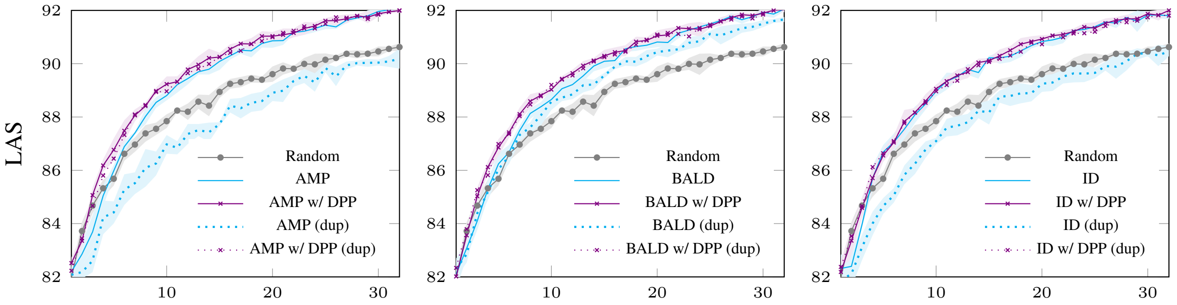

Figure 3: Learning curves of different sampling strategies based on AMP (left), BALD (middle) and ID (right), comparing diversity-aware (w/ DPP) and diversity-agnostic variants using the original and duplicated corpus (dup). The x-axis shows the number of rounds for annotation. Random (dup) curves overlap with those of Random and are omitted for readability.

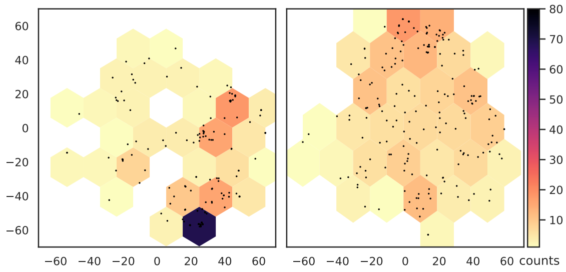

Figure 4: t-SNE visualization of the distributions of the 200 highest-quality unlabeled sentences over the diversity feature space after the 1st (left) and the 10th (right) annotation rounds using AMP without DPPs. Darker region indicates more data points residing in that diversity feature neighborhood. The left figure contains a dense region, while the data in the right figure are spread out in the feature space.

centrated within a small neighborhood. A diversity-agnostic strategy will sample similar examples for annotation. After a few rounds of annotation and model training, the distribution of high quality examples spreads out, and an AMP selection strategy is likely to sample a diverse set of examples without explicitly modeling diversity. Our analysis corroborates previous findings (Thompson et al., 1999) that small annotation batches are effective early in uncertainty sampling, avoiding selecting many near-duplicate examples when intra-batch diversity is low, but a larger batch size is more efficient later in training once intra-batch diversity increases.

5 Related Work

Modeling diversity in batch-mode AL (Brinker, 2003) has recently attracted attention in the machine learning community. Kirsch et al. (2019) introduce a Bayesian batch-mode selection strategy by estimating the mutual information between a set of samples and the model parameters. Ash et al. (2020) present a diversity-inducing sampling method using gradient embeddings. Most related to our work, Bıyık et al. (2019) first apply DPPs to batch-mode AL. Building on their approach, we flesh out a DPP treatment for AL for a structured prediction task, dependency parsing. Previously, Shen et al. (2018) consider named entity recognition but they report negative results for a diversity-inducing variant of their sampling method.

Due to the high annotation cost, AL is a popular technique for parsing and parse selection (Osborne and Baldridge, 2004). Recent advances focus on reducing full-sentence annotations to a subset of tokens within a sentence (Sassano and Kurohashi, 2010; Mirroshandel and Nasr, 2011; Majidi and Crane, 2013; Flannery and Mori, 2015; Li et al., 2016). We show that AL for parsing can further benefit from diversity-aware sampling strategies.

DPPs have previously been successfully applied to the tasks of extractive text summarization (Cho et al., 2019a,b) and modeling phoneme inventories (Cotterell and Eisner, 2017). In this work, we show that DPPs also provide a useful framework for understanding and modeling quality and diversity in active learning for NLP tasks.

6 Conclusion

We show that compared with their diversity-agnostic counterparts, diversity-aware sampling strategies not only lead to higher data efficiency, but are also more robust under corpus duplication settings. Our work invites future research into methods, utility and success conditions for modeling diversity in active learning for NLP tasks.

Acknowledgements

We thank the anonymous reviewers for their insightful reviews, and Prabhanjan Kambadur, Chen-Tse Tsai, and Minjie Xu for discussion and comments. Tianze Shi acknowledges support from Bloomberg's Data Science Ph.D. Fellowship.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, Online. OpenReview.net.

Ann Bies, Justin Mott, and Colin Warner. 2015. English news text treebank: Penn Treebank revised (LDC2015T13).

Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. 2019. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, Text, Speech and Language Technology, pages 103–127. Springer Netherlands, Dordrecht.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, pages 59–66, Washington, DC, USA. AAAI Press.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the Web. *Computer Networks and ISDN Systems*, 29(8):1157–1166.

Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5627–5638, Montréal, Canada.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019a. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

Sangwoo Cho, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2019b. Multi-document summarization with determinantal point processes and contextualized representations. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 98–103, Hong Kong, China. Association for Computational Linguistics.

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*, Toulon, France. OpenReview.net.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71B(4):233–240.

Daniel Flannery and Shinsuke Mori. 2015. Combining active learning and partial annotation for domain adaptation of a Japanese dependency parser. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 11–19, Bilbao, Spain. Association for Computational Linguistics.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1183–1192, Sydney, Australia. PMLR.

Rebecca Hwa. 2000. Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 45–52, Hong Kong, China. Association for Computational Linguistics.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30(3):253–276.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, San Diego, California, USA.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems 32*, pages 7026–7037, Vancouver, Canada. Curran Associates, Inc.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Terry Koo, Amir Globerson, Xavier Carreras, and Michael Collins. 2007. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150, Prague, Czech Republic. Association for Computational Linguistics.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2008. *Dependency Parsing*, volume 2 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286.

John A. Kulesza. 2012. *Learning with Determinantal Point Processes*. Ph.D. thesis, University of Pennsylvania.

David D. Lewis and William A. Gale. 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*, pages 3–12, Dublin, Ireland. ACM/Springer.

Zhenghua Li, Min Zhang, Yue Zhang, Zhanyi Liu, Wenliang Chen, Hua Wu, and Haifeng Wang. 2016. Active learning for dependency parsing with partial annotation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 344–354, Berlin, Germany.

Saeed Majidi and Gregory Crane. 2013. Active learning for dependency parsing by a committee of parsers. In *Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 98–105, Nara, Japan. Assocation for Computational Linguistics.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web*, pages 141–150, Banff, Alberta, Canada. Association for Computing Machinery.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Seyed Abolghasem Mirroshandel and Alexis Nasr. 2011. Active learning for dependency parsing using partially annotated sentences. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 140–149, Dublin, Ireland. Association for Computational Linguistics.

Miles Osborne and Jason Baldridge. 2004. Ensemble-based active learning for parse selection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 89–96, Boston, Massachusetts, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037, Vancouver, Canada. Curran Associates, Inc.

Manabu Sassano and Sadao Kurohashi. 2010. Using smaller constituents rather than sentences in active learning for Japanese dependency parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 356–365, Uppsala, Sweden. Association for Computational Linguistics.

Alexandra Schofield, Laure Thompson, and David Mimno. 2017. Quantifying the effects of text duplication on semantic models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2747, Copenhagen, Denmark. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*, Vancouver, Canada. OpenReview.net.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 120–127, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 406–414, Bled, Slovenia. Morgan Kaufmann Publishers Inc.

William T. Tutte. 1984. *Graph Theory*. Addison-Wesley Publishing Company.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Appendix A Dependency Parser

We adopt the deep biaffine dependency parser proposed by Dozat and Manning (2017). The parser is conceptually simple and yet competitive with state-of-the-art dependency parsers. The parser has three components: feature extraction, unlabeled parsing and relation labeler.

Feature Extraction For a length-n sentence $x = x_0, x_1, x_2, \ldots, x_n$, where x_0 is the dummy root symbol, we extract contextualized features at each word position. In our experiments, we use a pre-trained multilingual XLM-RoBERTa base model (Conneau et al., 2020), and fine-tune the feature extractor along with the rest of our parser:

$$[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n] = \text{XLM-R}(x_0, x_1, \ldots, x_n).$$

Each word input to the XLM-RoBERTa model is processed with the SentencePiece tokenizer (Kudo and Richardson, 2018), and we follow Kitaev et al. (2019) and retain the vectors corresponding to the last sub-word units as their representations. For $\mathbf{x}_0$, we use the vector of the `[CLS]` token, which is appended by XLM-RoBERTa to the beginning of each sentence.

Unlabeled Parser The parser uses a deep biaffine attention mechanism to derive locally-normalized attachment probabilities for all potential head-dependent pairs:

$$\begin{aligned}
\mathbf{h}_i^{\text{arc-head}} &= \text{MLP}^{\text{arc-head}}(\mathbf{x}_i) \\
\mathbf{h}_j^{\text{arc-dep}} &= \text{MLP}^{\text{arc-dep}}(\mathbf{x}_j) \\
s_{i,j} &= [\mathbf{h}_i^{\text{arc-head}}; 1]^\top U^{\text{arc}} [\mathbf{h}_j^{\text{arc-dep}}; 1] \\
P_{\text{att}}(\text{head(j)} = i \mid x) &= \text{softmax}_i(s_{:,j}),
\end{aligned}$$

where $\text{MLP}^{\text{arc-head}}$ and $\text{MLP}^{\text{arc-dep}}$ are two multi-layer perceptrons (MLPs) projecting $\mathbf{x}$ vectors into d^{arc}-dimensional $\mathbf{h}$ vectors, $[;1]$ appends an element of 1 at the end of the vectors, and $U^{\text{arc}} \in \mathbb{R}^{(d^{\text{arc}}+1)\times(d^{\text{arc}}+1)}$ is a bilinear scoring matrix. This component is trained with cross-entropy loss of the gold-standard attachments. During inference, we use the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to find the spanning tree with the highest product of locally-normalized attachment probabilities.

Relation Labeler The relation labeling component employs a similar deep biaffine scoring function as the unlabeled parsing component:

$$\begin{aligned}
\mathbf{h}_i^{\text{rel-head}} &= \text{MLP}^{\text{rel-head}}(\mathbf{x}_i) \\
\mathbf{h}_j^{\text{rel-dep}} &= \text{MLP}^{\text{rel-dep}}(\mathbf{x}_j) \\
t_{i,j,r} &= [\mathbf{h}_i^{\text{rel-head}}; 1]^\top U_r^{\text{rel}} [\mathbf{h}_j^{\text{rel-dep}}; 1] \\
P(\text{rel}(i,j) = r) &= \text{softmax}_r(t_{i,j,:}),
\end{aligned}$$

where each $U_r^{\text{rel}} \in \mathbb{R}^{(d^{\text{rel}}+1)\times(d^{\text{rel}}+1)}$, and there are as many such matrices as the size of the dependency relation label set $|R|$. The relation labeler is trained using cross entropy loss on the gold-standard head-dependent pairs. During inference, the labeling decision for each arc is made independently given the predicted unlabeled parse tree.

Round #	5		10	
Strategy	w/o DPP	w/ DPP	w/o DPP	w/ DPP
Random	$89.01_{\pm.28}$	$\mathbf{89.67}_{\pm.30}$	$90.78_{\pm.27}$	$\mathbf{91.22}_{\pm.22}$
AMP	$89.67_{\pm.29}$	$\mathbf{90.24}_{\pm.39}$	$92.03_{\pm.10}$	$\mathbf{92.17}_{\pm.22}$
BALD	$89.82_{\pm.36}$	$\mathbf{90.29}_{\pm.20}$	$91.87_{\pm.36}$	$\mathbf{92.00}_{\pm.08}$
ID	$\mathbf{90.24}_{\pm.20}$	$90.03_{\pm.18}$	$\mathbf{92.16}_{\pm.17}$	$92.06_{\pm.16}$

Table B1: UAS after 5 and 10 rounds of annotation (roughly 5,000 and 7,000 training tokens respectively), comparing strategies with and without modeling diversity through DPP.

Appendix B Results with UAS Evaluation

We also evaluate different learning strategies based on unlabeled attachment scores (UAS), and the results are shown in Table B1. In line with LAS-based experiments, we find that modeling diversity is more helpful during initial stages of learning. For ID, we observe this effect even earlier than the fifth round of annotation: $86.57_{\pm.44}$ vs. $87.40_{\pm.51}$ after the first annotation round.

Appendix C Sentence Selection Examples

In Table C2 we compare batches sampled by a diversity-aware selection strategy with a diversity-agnostic one. We observe that by modeling diversity in the sample selection process, DPPs avoid selecting duplicate or near-duplicate sentences and thus the annotation budget can be maximally utilized.

Appendix D BALD under High Duplication Setting

Figure D1 shows the learning curves for BALD-based selection strategies under a high corpus duplication setting where the corpus is repeated

	Sentences selected by AMP (highest-quality ones first):
	Downgraded by Moody 's were Houston Lighting 's first - mortgage bonds and secured pollution - control bonds to single - A - 3 from single - A - 2 ; unsecured pollution - control bonds to Baa - 1 from single - A - 3 ; preferred stock to single - A - 3 from single - A - 2 ; a shelf registration for preferred stock to a preliminary rating of single - A - 3 from a preliminary rating of single - A - 2 ; two shelf registrations for collateralized debt securities to a preliminary rating of single - A - 3 from a preliminary rating of single - A - 2 , and the unit 's rating for commercial paper to Prime - 2 from Prime - 1 .
	For a while in the 1970s it seemed Mr. Moon was on a spending spree , with such purchases as the former New Yorker Hotel and its adjacent Manhattan Center ; a fishing / processing conglomerate with branches in Alaska , Massachusetts , Virginia and Louisiana ; a former Christian Brothers monastery and the Seagram family mansion (both picturesquely situated on the Hudson River) ; shares in banks from Washington to Uruguay ; a motion picture production company , and newspapers , such as the Washington Times , the New York City Tribune (originally the News World) , and the successful Spanish - language Noticias del Mundo .
→	LONDON LATE EURODOLLARS : 8 11/16 % to 8 9/16 % one month ; 8 5/8 % to 8 1/2 % two months ; 8 5/8 % to 8 1/2 % three months ; 8 9/16 % to 8 7/16 % four months ; 8 1/2 % to 8 3/8 % five months ; 8 1/2 % to 8 3/8 % six months .
→	LONDON LATE EURODOLLARS : 8 3/4 % to 8 5/8 % one month ; 8 3/4 % to 8 5/8 % two months ; 8 11/16 % to 8 9/16 % three months ; 8 9/16 % to 8 7/16 % four months ; 8 1/2 % to 8 3/8 % five months ; 8 7/16 % to 8 5/16 % six months .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.40 % 30 to 44 days ; 8.325 % 45 to 59 days ; 8.10 % 60 to 89 days ; 8 % 90 to 119 days ; 7.85 % 120 to 149 days ; 7.70 % 150 to 179 days ; 7.375 % 180 to 270 days .
	4 . When a RICO TRO is being sought , the prosecutor is required , at the earliest appropriate time , to state publicly that the government 's request for a TRO , and eventual forfeiture , is made in full recognition of the rights of third parties – that is , in requesting the TRO , the government will not seek to disrupt the normal , legitimate business activities of the defendant ; will not seek through use of the relation - back doctrine to take from third parties assets legitimately transferred to them ; will not seek to vitiate legitimate business transactions occurring between the defendant and third parties ; and will , in all other respects , assist the court in ensuring that the rights of third parties are protected , through proceeding under RICO and otherwise .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.50 % 30 to 44 days ; 8.25 % 45 to 62 days ; 8.375 % 63 to 89 days ; 8 % 90 to 119 days ; 7.90 % 120 to 149 days ; 7.80 % 150 to 179 days ; 7.55 % 180 to 270 days .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.50 % 30 to 44 days ; 8.25 % 45 to 65 days ; 8.375 % 66 to 89 days ; 8 % 90 to 119 days ; 7.875 % 120 to 149 days ; 7.75 % 150 to 179 days ; 7.50 % 180 to 270 days .
→	LONDON LATE EURODOLLARS : 8 11/16 % to 8 9/16 % one month ; 8 5/8 % to 8 1/2 % two months ; 8 5/8 % to 8 1/2 % three months ; 8 9/16 % to 8 7/16 % four months ; 8 1/2 % to 8 3/8 % five months ; 8 7/16 % to 8 5/16 % six months .
→	LONDON LATE EURODOLLARS : 8 11/16 % to 8 9/16 % one month ; 8 9/16 % to 8 7/16 % two months ; 8 5/8 % to 8 1/2 % three months ; 8 1/2 % to 8 3/8 % four months ; 8 7/16 % to 8 5/16 % five months ; 8 7/16 % to 8 5/16 % six months .
	The new edition lists the top 10 metropolitan areas as Anaheim - Santa Ana , Calif. ; Boston ; Louisville , Ky. ; Nassau - Suffolk , N.Y. ; New York ; Pittsburgh ; San Diego ; San Francisco ; Seattle ; and Washington .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.45 % 30 to 44 days ; 8.20 % 45 to 67 days ; 8.325 % 68 to 89 days ; 8 % 90 to 119 days ; 7.875 % 120 to 149 days ; 7.75 % 150 to 179 days ; 7.50 % 180 to 270 days .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.50 % 2 to 44 days ; 8.25 % 45 to 69 days ; 8.40 % 70 to 89 days ; 8.20 % 90 to 119 days ; 8.05 % 120 to 149 days ; 7.90 % 150 to 179 days ; 7.50 % 180 to 270 days .
	Five officials of this investment banking firm were elected directors : E. Garrett Bewkes III , a 38 - year - old managing director in the mergers and acquisitions department ; Michael R. Dabney , 44 , a managing director who directs the principal activities group which provides funding for leveraged acquisitions ; Richard Harriton , 53 , a general partner who heads the correspondent clearing services ; Michael Minikes , 46 , a general partner who is treasurer ; and William J. Montgoris , 42 , a general partner who is also senior vice president of finance and chief financial officer .
→	LONDON LATE EURODOLLARS : 8 11/16 % to 8 9/16 % one month ; 8 5/8 % to 8 1/2 % two months ; 8 11/16 % to 8 9/16 % three months ; 8 9/16 % to 8 7/16 % four months ; 8 1/2 % to 8 3/8 % five months ; 8 7/16 % to 8 5/16 % six months .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.55 % 30 to 44 days ; 8.25 % 45 to 59 days ; 8.40 % 60 to 89 days ; 8 % 90 to 119 days ; 7.90 % 120 to 149 days ; 7.80 % 150 to 179 days ; 7.55 % 180 to 270 days .
	They transferred some $ 28 million from the Community Development Block Grant program designated largely for low - and moderate - income projects and funneled it into such items as : – $ 1.2 million for a performing - arts center in Newark , – $ 1.3 million for " job retention " in Hawaiian sugar mills . – $ 400,000 for a collapsing utility tunnel in Salisbury , – $ 500,000 for " equipment and landscaping to deter crime and aid police surveillance " at a Michigan park . – $ 450,000 for " integrated urban data based in seven cities . " No other details . – $ 390,000 for a library and recreation center at Mackinac Island , Mich .
→	LONDON LATE EURODOLLARS : 8 3/4 % to 8 5/8 % one month ; 8 13/16 % to 8 11/16 % two months ; 8 11/16 % to 8 9/16 % three months ; 8 9/16 % to 8 7/16 % four months ; 8 1/2 % to 8 3/8 % five months ; 8 7/16 % to 8 5/16 % six months .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.45 % 30 to 44 days ; 8.25 % 45 to 68 days ; 8.30 % 69 to 89 days ; 8.125 % 90 to 119 days ; 8 % 120 to 149 days ; 7.875 % 150 to 179 days ; 7.50 % 180 to 270 days .
	Sentences selected by AMP with diversity-inducing DPP:
	Downgraded by Moody 's were Houston Lighting 's first - mortgage bonds and secured pollution - control bonds to single - A - 3 from single - A - 2 ; unsecured pollution - control bonds to Baa - 1 from single - A - 3 ; preferred stock to single - A - 3 from single - A - 2 ; a shelf registration for preferred stock to a preliminary rating of single - A - 3 from a preliminary rating of single - A - 2 ; two shelf registrations for collateralized debt securities to a preliminary rating of single - A - 3 from a preliminary rating of single - A - 2 , and the unit 's rating for commercial paper to Prime - 2 from Prime - 1 .
	4 . When a RICO TRO is being sought , the prosecutor is required , at the earliest appropriate time , to state publicly that the government 's request for a TRO , and eventual forfeiture , is made in full recognition of the rights of third parties – that is , in requesting the TRO , the government will not seek to disrupt the normal , legitimate business activities of the defendant ; will not seek through use of the relation - back doctrine to take from third parties assets legitimately transferred to them ; will not seek to vitiate legitimate business transactions occurring between the defendant and third parties ; and will , in all other respects , assist the court in ensuring that the rights of third parties are protected , through proceeding under RICO and otherwise .
▷	COMMERCIAL PAPER placed directly by General Motors Acceptance Corp. : 8.40 % 30 to 44 days ; 8.325 % 45 to 59 days ; 8.10 % 60 to 89 days ; 8 % 90 to 119 days ; 7.85 % 120 to 149 days ; 7.70 % 150 to 179 days ; 7.375 % 180 to 270 days .
	Moreover , the process is n't without its headaches .
	For a while in the 1970s it seemed Mr. Moon was on a spending spree , with such purchases as the former New Yorker Hotel and its adjacent Manhattan Center ; a fishing / processing conglomerate with branches in Alaska , Massachusetts , Virginia and Louisiana ; a former Christian Brothers monastery and the Seagram family mansion (both picturesquely situated on the Hudson River) ; shares in banks from Washington to Uruguay ; a motion picture production company , and newspapers , such as the Washington Times , the New York City Tribune (originally the News World) , and the successful Spanish - language Noticias del Mundo .
	Within the paper sector , Mead climbed 2 3/8 to 38 3/4 on 1.3 million shares , Union Camp rose 2 3/4 to 37 3/4 , Federal Paper Board added 1 3/4 to 23 7/8 , Bowater gained 1 1/2 to 27 1/2 , Stone Container rose 1 to 26 1/8 and Temple - Inland jumped 3 3/4 to 62 1/4 .
	We finally rendezvoused with our balloon , which had come to rest on a dirt road amid a clutch of Epinalers who watched us disassemble our craft – another half - an - hour of non-flight activity – that included the precision routine of yanking the balloon to the ground , punching all the air out of it , rolling it up and cramming it and the basket into the trailer .
	These are the 26 states , including the commonwealth of Puerto Rico , that have settled with Drexel : Alaska , Arkansas , Delaware , Georgia , Hawaii , Idaho , Indiana , Iowa , Kansas , Kentucky , Maine , Maryland , Minnesota , Mississippi , New Hampshire , New Mexico , North Dakota , Oklahoma , Oregon , South Carolina , South Dakota , Utah , Vermont , Washington , Wyoming and Puerto Rico .
	It is the stuff of dreams , but also of traumas .
	An inquiry into his handling of Lincoln S&L inevitably will drag in Sen. Cranston and the four others , Sens. Dennis DeConcini (D. , Ariz.) , John McCain (R. , Ariz.) , John Glenn (D. , Ohio) and Donald Riegle (D. , Mich .) .
	Five officials of this investment banking firm were elected directors : E. Garrett Bewkes III , a 38 - year - old managing director in the mergers and acquisitions department ; Michael R. Dabney , 44 , a managing director who directs the principal activities group which provides funding for leveraged acquisitions ; Richard Harriton , 53 , a general partner who heads the correspondent clearing services ; Michael Minikes , 46 , a general partner who is treasurer ; and William J. Montgoris , 42 , a general partner who is also senior vice president of finance and chief financial officer .
	But as they hurl fireballs that smolder rather than burn , and relive old duels in the sun , it 's clear that most are there to make their fans cheer again or recapture the camaraderie of seasons past or prove to themselves and their colleagues that they still have it – or something close to it .
	They are : " A Payroll to Meet : A Story of Greed , Corruption and Football at SMU " (Macmillan , 221 pages , $ 18.95) by David Whitford ; " Big Red Confidential : Inside Nebraska Football " (Contemporary , 231 pages , $ 17.95) by Armen Keteyian ; and " Never Too Young to Die : The Death of Len Bias " (Pantheon , 252 pages , $ 18.95) by Lewis Cole .
	He says he told NewsEdge to look for stories containing such words as takeover , acquisition , acquire , LBO , tender , merger , junk and halted .
	It is no coincidence that from 1844 to 1914 , when the Bank of England was an independent private bank , the pound was never devalued and payment of gold for pound notes was never suspended , but with the subsequent nationalization of the Bank of England , the pound was devalued with increasing frequency and its use as an international medium of exchange declined .
	The $ 4 billion in bonds break down as follows : $ 1 billion in five - year bonds with a coupon rate of 8.25 % and a yield to maturity of 8.33 % ; $ 1 billion in 10 - year bonds with a coupon rate of 8.375 % and a yield to maturity of 8.42 % ; $ 2 billion in 30 - year bonds with five - year call protection , a coupon rate of 8.75 % and a yield to maturity of 9.06 % .
	Hecla Mining rose 5/8 to 14 ; Battle Mountain Gold climbed 3/4 to 16 3/4 ; Homestake Mining rose 1 1/8 to 16 7/8 ; Lac Minerals added 5/8 to 11 ; Placer Dome went up 7/8 to 16 3/4 , and ASA Ltd. jumped 3 5/8 to 49 5/8 .

Table C2: Sentences picked by a diversity-agnostic (top) and a diversity-aware (bottom) selection strategy from the same unlabeled pool after the intial round of model training on the seed sentences. The diversity-agnostic strategy selects many near-duplicate sentences (the two near-duplicate clusters are marked by red → and blue ▷), effectively wasting the annotation budget, where DPPs largely alleviate this issue by enforcing diversity.

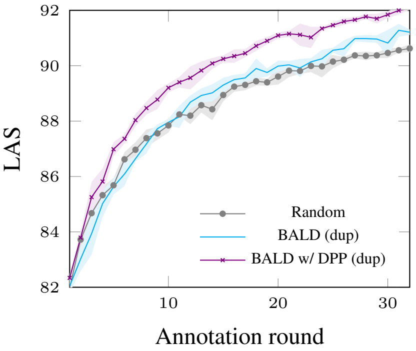

Figure D1: Learning curves for BALD-based selection strategies under a five-fold corpus duplication setting.

Algorithm E1: Greedy MAP inference for DPP with a size budget, adapted from Kulesza (2012).

```
Input: candidate item set $X$ (sentences or tokens),
       DPP represented by matrix $L$, size budget $b$
$U \leftarrow X$;
$Y \leftarrow \emptyset$;
while $U \neq \emptyset$ do
    $i \leftarrow \arg\max_{i' \in U} \det(L_{Y \cup \{i'\}})$;
    if $\sum_{y \in Y} \text{size}(y) < b$ then
        $Y \leftarrow Y \cup \{i\}$;
    else
        break;
    end
end
Output: selected items $Y$
```

five times. In this extreme setting, the diversity-agnostic strategy significantly underperforms the diversity-aware one. We posit that the relative success of BALD compared to AMP in the twice-duplicated setting is due to the fact that BALD randomly draws dropout masks to estimate model uncertainty, so that identical examples could still have different quality measures.

Appendix E Implementation Details and Hyperparameters

We do not tune our hyperparameters since in practice, active learning systems only have a single shot at success, without tuning. Instead, we follow recommendations from relevant prior work in setting our learning details and hyperparameters.

Active Learning Following Li et al. (2016), our active learning set-up proceeds in two stages for each annotation round. In the first stage, we select sentences filling in a budget of 2500 tokens; in the second stage, we pick 500 tokens out of the subset of sentences. For a diversity-agnostic strategy, we choose the top-k highest-quality candidates within the token budget, while our diversity-aware selection strategy uses a separate DPP for each stage. The active learning process is bootstraped with a seed set of 128 labeled sentences. For the BALD quality measure, we set $K = 5$.

Greedy MAP Inference for DPPs Algorithm E1 illustrates the procedure for selecting items from DPPs under a budget constraint. This greedy MAP inference algorithm is adapted from Kulesza (2012). During sentence selection, the size of a sentence is its number of tokens, and each token has a size of 1 in the token selection stage.

Dependency Parser We set the hyperparameters according to Dozat and Manning (2017). All the MLPs in the deep biaffine attention architecture have single hidden layers with ReLU activation functions and a dropout probability of 0.33, and we set d^{arc} and d^{rel} to be 500 and 100 respectively.

Training and Optimization Each training batch contains 16 sentences and gradient norms are clipped to 5.0. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-5} with 640 warmup steps with a linearly-increasing learning rate starting from 0.

Implementation Our implementation is in PyTorch (Paszke et al., 2019), and we use the `transformers` package[6] to interface with the pre-trained XLM-RoBERTa model.

[6] https://github.com/huggingface/transformers