Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers

André F. T. Martins*†
Miguel B. Almeida*†
Noah A. Smith#

*Priberam Labs, Alameda D. Afonso Henrique, 41, 2º, 1000-123 Lisboa, Portugal
†Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
#School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

{atm,mba}@priberam.pt, nasmith@cs.cmu.edu

Abstract

We present fast, accurate, direct non-projective dependency parsers with third-order features. Our approach uses AD³, an accelerated dual decomposition algorithm which we extend to handle specialized head automata and sequential head bigram models. Experiments in fourteen languages yield parsing speeds competitive to projective parsers, with state-of-the-art accuracies for the largest datasets (English, Czech, and German).

1 Introduction

Dependency parsing has become a prominent approach to syntax in the last few years, with increasingly fast and accurate models being devised (Kübler et al., 2009; Huang and Sagae, 2010; Zhang and Nivre, 2011; Rush and Petrov, 2012).

In projective parsing, the arcs in the dependency tree are constrained to be nested, and the problem of finding the best tree can be addressed with dynamic programming. This results in cubic-time decoders for arc-factored and sibling second-order models (Eisner, 1996; McDonald and Pereira, 2006), and quartic-time for grandparent models (Carreras, 2007) and third-order models (Koo and Collins, 2010). Recently, Rush and Petrov (2012) trained third-order parsers with vine pruning cascades, achieving runtimes only a small factor slower than first-order systems. Third-order features have also been included in transition systems (Zhang and Nivre, 2011) and graph-based parsers with cube-pruning (Zhang and McDonald, 2012).

Unfortunately, non-projective dependency parsers (appropriate for languages with a more flexible word order, such as Czech, Dutch, and German) lag behind these recent advances. The main obstacle is that non-projective parsing is NP-hard beyond arc-factored models (McDonald and Satta, 2007). Approximate parsers have therefore been introduced, based on belief propagation (Smith and Eisner, 2008), dual decomposition (Koo et al., 2010), or multi-commodity flows (Martins et al., 2009, 2011). These are all instances of turbo parsers, as shown by Martins et al. (2010): the underlying approximations come from the fact that they run global inference in factor graphs ignoring loop effects. While this line of research has led to accuracy gains, none of these parsers use third-order contexts, and their speeds are well behind those of projective parsers.

This paper bridges the gap above by presenting the following contributions:

- We apply the third-order feature models of Koo and Collins (2010) to non-projective parsing.
- This extension is non-trivial since exact dynamic programming is not applicable. Instead, we adapt AD³, the dual decomposition algorithm proposed by Martins et al. (2011), to handle third-order features, by introducing specialized head automata.
- We make our parser substantially faster than the many-components approach of Martins et al. (2011). While AD³ requires solving quadratic subproblems as an intermediate step, recent results (Martins et al., 2012) show that they can be addressed with the same oracles used in the subgradient method (Koo et al., 2010). This enables AD³ to exploit combinatorial subproblems like the head automata above.

Along with this paper, we provide a free distribution of our parsers, including training code.
Given a sentence with \( L \) words, to which we prepend a root symbol $\$, let \( A := \{ (h, m) \mid h \in \{0, \ldots, L\}, m \in \{1, \ldots, L\}, h \neq m \} \) be the set of possible dependency arcs. We parameterize a dependency tree via an indicator vector \( u := (u_a)_{a \in A} \), where \( u_a = 1 \) if the arc \( a \) is in the tree, and 0 otherwise, and we denote by \( \mathcal{Y} \subseteq [1, \ldots, S] \) the set of such vectors that are indicators of well-formed trees. Let \( \{ A_s \}_{s=1}^S \) be a cover of \( A \), where each \( A_s \subseteq A \). We assume that the score of a parse tree \( u \in \mathcal{Y} \) decomposes as \( f(u) := \sum_{s=1}^S f_s(z_s) \), where each \( z_{s,a} := (z_{s,a})_{a \in A_s} \) is a “partial view” of \( u \), and each local score function \( f_s \) comes from a feature-based linear model.

Past work in dependency parsing considered either (i) a few “large” components, such as trees and head automata (Smith and Eisner, 2008; Koo et al., 2010), or (ii) many “small” components, coming from a multi-commodity flow formulation (Martins et al., 2009, 2011). Let \( \mathcal{Y}_s \subseteq \mathbb{R}^{|A_s|} \) denote the set of feasible realizations of \( z_s \), i.e., those that are partial views of an actual parse tree. A tuple of views \( \langle z_1, \ldots, z_S \rangle \in \prod_{s=1}^S \mathcal{Y}_s \) is said to be globally consistent if \( z_{s,a} = z_{sIT}, a \) holds for every \( a, s, \) and \( s' \) such that \( a \in A_s \cap A_{s'} \). We assume each parse \( u \in \mathcal{Y} \) corresponds uniquely to a globally consistent tuple of views, and vice-versa. Following Martins et al. (2011), the problem of obtaining the best-scored tree can be written as follows:

\[
\text{maximize } \sum_{s=1}^S f_s(z_s) \\
\text{w.r.t. } u \in \mathbb{R}^{|A|}, \quad z_{s,a} = u_a, \quad \forall s, \quad \forall a \in A_s, \quad (1)
\]

where the equality constraint ensures that the partial views “glue” together to form a coherent parse tree.\(^3\)

### 2.2 Dual Decomposition and AD\(^3\)

Dual decomposition methods dualize out the equality constraint in Eq. 1 by introducing Lagrange multipliers \( \lambda_{s,a} \). In doing so, they solve a relaxation where the combinatorial sets \( \mathcal{Y}_s \) are replaced by their convex hulls \( \mathcal{Z}_s := \text{conv}(\mathcal{Y}_s) \). All that is necessary is the following assumption:

**Assumption 1** (Local-Max Oracle). Every \( s \in \{1, \ldots, S\} \) has an oracle that solves efficiently any instance of the following subproblem:

\[
\text{maximize } f_s(z_s) + \sum_{a \in A_s} \lambda_{s,a} z_{s,a} \\
\text{w.r.t. } z_s \in \mathcal{Z}_s. \quad (2)
\]

Typically, Assumption 1 is met whenever the maximization of \( f_s \) over \( \mathcal{Z}_s \) is tractable, since the objective in Eq. 2 just adds a linear function to \( f_s \).

---

\( ^2 \)Concretely, AD\(^3\) needs \( O(1/\epsilon) \) iterations to converge to a \( \epsilon \)-accurate solution, while subgradient needs \( O(1/\epsilon^2) \).

\( ^3 \)Note that any tuple \( \langle z_1, \ldots, z_S \rangle \in \prod_{s=1}^S \mathcal{Y}_s \) satisfying the equality constraints will be globally consistent; this fact, due the assumptions above, will imply \( u \in \mathcal{Y} \).

\( ^4 \)Let \( \Delta^{\mathcal{Y}_s} := \{ \alpha \in \mathbb{R}^{|\mathcal{Y}_s|} \mid \alpha \geq 0, \sum_{y \in \mathcal{Y}_s} \alpha_y = 1 \} \) be the probability simplex. The convex hull of \( \mathcal{Y}_s \) is the set \( \text{conv}(\mathcal{Y}_s) := \{ \sum_{y \in \mathcal{Y}_s} \alpha_y y \mid \alpha \in \Delta^{\mathcal{Y}_s} \} \). Its members represent marginal probabilities over the arcs in \( A_s \).
Proposition 2. The problem in Eq. 3 admits a solution $z^*_s$ which is spanned by a sparse basis $W \subseteq Y_s$ with cardinality at most $|W| \leq O(|A_s|)$. In other words, there is a distribution $\alpha$ with support in $W$ such that $z^*_s = \sum_{y_s \in W} y_s, \alpha y_s$.\footnote{In our experiments (§4), we set $\rho = 0.05.$}

Prop. 2 has motivated an active set algorithm (Martins et al., 2012) that maintains an estimate of $W$ by iteratively adding and removing elements computed through the oracle in Eq. 2.\footnote{Note that $|y_s| = O(2^{|A_s|})$ in general. What Prop. 2 tells us is that the solution of Eq. 3 can be represented as a distribution over $y_s$, with a very sparse support.} Typically, very few iterations are necessary and great speed-ups are achieved by warm-starting $W$ with the active set computed in the previous AD$^3$ iteration. This has a huge impact in practice and is crucial to obtain the fast runtimes in §4 (see Fig. 2).

The AD$^3$ algorithm (Martins et al., 2011) alternates among the following iterative updates:

- **$z$-updates**, which decouple over $s = 1, \ldots, S$, and solve a penalized version of Eq. 2:
  \[
  z^{(t+1)}_s := \text{arg max}_{z_s \in Z_s} f_s(z_s) + \sum_{a \in A_s} \lambda^{(t)}_{s,a} z_{s,a} - \frac{\rho}{2} \sum_{a \in A_s} (z_{s,a} - u^{(t)}_a)^2. \tag{3}
  \]
  Above, $\rho$ is a constant and the quadratic term penalizes deviations from the current global solution (stored in $u^{(t)}$).\footnote{The algorithm is a specialization of Nocedal and Wright (1999), §16.4, which effectively exploits the sparse representation of $z$. For details, see Martins et al. (2012).} We will see (Prop. 2) that this problem can be solved iteratively using only the Local-Max Oracle (Eq. 2).

- **$u$-updates**, a simple averaging operation:
  \[
  u^{(t+1)}_a := \frac{1}{|\{s : a \in A_s\}|} \sum_{s : a \in A_s} z^{(t+1)}_s. \tag{4}
  \]

- **$\lambda$-updates**, where the Lagrange multipliers are adjusted to penalize disagreements:
  \[
  \lambda^{(t+1)}_{s,a} := \lambda^{(t)}_{s,a} - \rho (z^{(t+1)}_{s,a} - u^{(t+1)}_a). \tag{5}
  \]

3 Solving the Subproblems

We next describe the actual components used in our third-order parsers.

**Tree component.** We use an arc-factored score function (McDonald et al., 2005):
\[
  f_{\text{TREE}}(z) = \sum_{m=1}^{|L|} \sigma_{\text{ARC}}(\pi(m), m),
\]
where $\pi(m)$ is the parent of the $m$th word according to the parse tree $z$, and $\sigma_{\text{ARC}}(h, m)$ is the score of an individual arc. The parse tree that maximizes this function can be found in time $O(L^3)$ via the Chu-Liu-Edmonds’ algorithm (Chu and Liu, 1965; Edmonds, 1967).\footnote{In fact, there is an asymptotically faster $O(L^2)$ algorithm (Tarjan, 1977). Moreover, if the set of possible arcs is reduced to a subset $B \subseteq A$ (via pruning), then the fastest known algorithm (Gabow et al., 1986) runs in $O(|B| + L \log L)$ time.}

**Grand-sibling head automata.** Let $A^\text{in}_h$ and $A^\text{out}_h$ denote respectively the sets of incoming and outgoing candidate arcs for the $h$th word, where the latter subdivides into arcs pointing to the right, $A^\text{out}_h \rightarrow$, and to the left, $A^\text{out}_h \leftarrow$. Define the sets $A^\text{GSIB}_h = A^\text{in}_h \cup A^\text{out}_h \rightarrow$ and $A^\text{GSIB}_h = A^\text{in}_h \cup A^\text{out}_h \leftarrow$. We describe right-side grand-sibling head automata; their left-side counterparts are analogous. For each head word $h$ in the parse tree $z$, define $g := \pi(h)$, and let $\langle m_0, m_1, \ldots, m_{p+1} \rangle$ be the sequence of right modifiers of $h$, with $m_0 = \text{START}$ and $m_{p+1} = \text{END}$. Then, we have the following grand-sibling component:
\[
f_{\text{GSIB}}(z|A^\text{GSIB}_h) = \sum_{k=1}^{p+1} \left( \sigma_{\text{SIB}}(h, m_{k-1}, m_k) \right.
\]
\[
+ \sigma_{\text{GP}}(g, h, m_k) + \sigma_{\text{GSIB}}(g, h, m_{k-1}, m_k) \bigg),
\]
where we use the shorthand $z|B$ to denote the subvector of $z$ indexed by the arcs in $B \subseteq A$. Note that this score function absorbs grandparent and consecutive sibling scores, in addition to the grand-sibling scores.\footnote{Koo et al. (2010) used an identical automaton for their second-order model, but leaving out the grand-sibling scores.} For each $h$, $f_{\text{GSIB}}^{A^\text{GSIB}_h}$ can be found in time $O(|A|)$ by using the algorithm of Chu and Liu (1965) with pruning.

Figure 2: Comparison between AD$^3$ and subgradient. We show averaged runtimes in PTB §22 as a function of the sentence length. For subgradient, we chose for each sentence the most favorable stepsize in $\{0.001, 0.01, 0.1, 1\}$.\footnote{In our experiments (§4), we set $\rho = 0.05$.}
maximized in time $O(L^3)$ with dynamic programming, yielding $O(L^3)$ total runtime.

**Tri-sibling head automata.** In addition, we define left and right-side tri-sibling head automata that remember the previous two modifiers of a head word. This corresponds to the following component function (for the right-side case):

$$f_{TSIB}^{\text{TSIB}}(z|_{A_{\text{out}}}) = \sum_{k=2}^{p+1} \sigma_{TSIB}(h, m_{k-2}, m_{k-1}, m_k).$$

Again, each of these functions can be maximized in time $O(L^3)$, yielding $O(L^3)$ runtime.

**Sequential head bigram model.** Head bigrams can be captured with a simple sequence model:

$$f_{\text{SEQ}}(z) = \sum_{m=2}^{L} \sigma_{\text{HB}}(m, \pi(m), \pi(m-1)).$$

Each score $\sigma_{\text{HB}}(m, h, h')$ is obtained via features that look at the heads of consecutive words (as in Martins et al. (2011)). This function can be maximized in time $O(L^3)$ with the Viterbi algorithm.

**Arbitrary siblings.** We handle arbitrary siblings as in Martins et al. (2011), defining $O(L^3)$ component functions of the form $f_{\text{ASIB}}^{\text{ASIB}}(z_{h,m,s}, z_{h,s}) = \sigma_{\text{ASIB}}(h, m, s)$. In this case, the quadratic problem in Eq. 3 can be solved directly in constant time.

Tab. 1 details the time complexities of each subproblem. Without pruning, each iteration of AD$^3$ has $O(L^4)$ runtime. With a simple strategy that limits the number of candidate heads per word to a constant $K$, this drops to cubic time.\(^{10}\) Further speed-ups are possible with more pruning: by limiting the number of possible modifiers to a constant $J$, the runtime would reduce to $O(L \log L)$.

\(^{10}\)In our experiments, we employed this strategy with $K = 10$, by pruning with a first-order probabilistic model. Following Koo and Collins (2010), for each word $m$, we also pruned away incoming arcs $(h, m)$ with posterior probability less than 0.0001 times the probability of the most likely head.

| TREE | $|A_m^n| \leq K$ same, $|A_m^m| \leq J$ | PTB-YM §22, 1st ord | PTB-YM §22, 2nd ord | PTB-YM §22, 2nd ord, +ASIB, +HB | PTB-YM §22, 3rd ord | PTB-YM §22, 3rd ord, gold tags |
|------|----------------------------------|-----------------|-----------------|-------------------------------|-----------------|-----------------|
| GSBIB | $O(L^2)$ | $O(KL + L \log L)$ | $O(KL + L \log L)$ | $O(LK + KL)$ | $O(JK^2L)$ | $O(JK^2L)$ |
| TSIB | $O(L^2)$ | $O(KL^2)$ | $O(J^2KL)$ | $O(J^2KL)$ | $O(J^2KL)$ | $O(J^2KL)$ |
| SEQ | $O(L^3)$ | $O(K^2L)$ | $O(K^2L)$ | $O(K^2L)$ | $O(K^2L)$ | $O(K^2L)$ |
| ASIB | $O(L^3)$ | $O(KL^2)$ | $O(JKL)$ | $O(JKL)$ | $O(JKL)$ | $O(JKL)$ |

**Table 1:** Theoretical runtimes of each subproblem without pruning, limiting the number of candidate heads, and limiting (in addition) the number of modifiers. Note the $O(L \log L)$ total runtime per AD$^3$ iteration in the latter case.

**4 Experiments**

We first evaluated our non-projective parser in a projective English dataset, to see how its speed and accuracy compares with recent projective parsers, which can take advantage of dynamic programming. To this end, we converted the Penn Treebank to dependencies through (i) the head rules of Yamada and Matsumoto (2003) (PTB-YM) and (ii) basic dependencies from the Stanford parser 2.0.5 (PTB-S).\(^{11}\) We trained by running 10 epochs of cost-augmented MIRA (Crammer et al., 2006). To ensure valid parse trees at test time, we rounded fractional solutions as in Martins et al. (2009)—yet, solutions were integral $\approx 95\%$ of the time.

Tab. 2 shows the results in the dev-set (top block) and in the test-set (two bottom blocks). In the dev-set, we see consistent gains when more expressive features are added, the best accuracies being achieved with the full third-order model; this comes at the cost of a 6-fold drop in runtime compared with a first-order model. By looking at the two bottom blocks, we observe that our parser has slightly better accuracies than recent projective parsers, with comparable speed levels (with the exception of the highly optimized vine cascade approach of Rush and Petrov, 2012).

\(^{11}\)We train on sections §02–21, use §22 as validation data, and test on §23. We trained a simple 2nd-order tagger with 10-fold jackknifing to obtain automatic part-of-speech tags for §22–23, with accuracies 97.2% and 90.9%, respectively.
Table 3: Results for the CoNLL-2006 datasets and the non-projective English dataset of CoNLL-2008.

<table>
<thead>
<tr>
<th>Language</th>
<th>First Ord. UAS</th>
<th>Sec. Ord. UAS</th>
<th>Third Ord. UAS</th>
<th>Best published UAS</th>
<th>RP12 UAS</th>
<th>ZM12 UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>77.23 2,481</td>
<td>78.50 2,481</td>
<td>79.64 252</td>
<td>81.12 3,980</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>91.76 5,678</td>
<td>92.82 2,049</td>
<td>93.10 1,273</td>
<td>93.50 3,980</td>
<td>91.9</td>
<td>93.08</td>
</tr>
<tr>
<td>Chinese</td>
<td>88.49 18,094</td>
<td>90.14 4,284</td>
<td>91.89 2,592</td>
<td>92.30 8,600</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Czech</td>
<td>87.66 1,840</td>
<td>90.00 751</td>
<td>90.32 751</td>
<td>90.46 2,880</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Danish</td>
<td>89.42 4,110</td>
<td>91.20 1,053</td>
<td>91.48 650</td>
<td>91.86 3,980</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dutch</td>
<td>83.61 3,884</td>
<td>86.37 1,294</td>
<td>86.19 599</td>
<td>85.81 121 Ko10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>German</td>
<td>90.52 5,331</td>
<td>91.85 1,788</td>
<td>92.41 965</td>
<td>91.89 2,880</td>
<td>90.8</td>
<td>91.35</td>
</tr>
<tr>
<td>English</td>
<td>91.21 3,127</td>
<td>93.03 1,317</td>
<td>93.22 785</td>
<td>92.68 2,880</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Japanese</td>
<td>92.78 23,895</td>
<td>93.14 5,660</td>
<td>93.52 2,996</td>
<td>93.72 8,600</td>
<td>92.3</td>
<td>93.24</td>
</tr>
<tr>
<td>Portuguese</td>
<td>91.14 4,273</td>
<td>92.71 1,316</td>
<td>92.99 740</td>
<td>93.03 79 Ko10</td>
<td>91.5</td>
<td>91.69</td>
</tr>
<tr>
<td>Slovene</td>
<td>82.81 4,315</td>
<td>85.21 722</td>
<td>86.01 366</td>
<td>86.95 2,880</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Spanish</td>
<td>83.61 4,347</td>
<td>84.97 623</td>
<td>85.59 318</td>
<td>87.48 2,880</td>
<td>90.1</td>
<td>91.44</td>
</tr>
<tr>
<td>Swedish</td>
<td>89.36 5,622</td>
<td>90.98 1,387</td>
<td>91.14 684</td>
<td>91.44 2,880</td>
<td>90.1</td>
<td>91.44</td>
</tr>
<tr>
<td>Turkish</td>
<td>75.98 6,418</td>
<td>76.50 1,721</td>
<td>76.90 793</td>
<td>77.55 258 Ko10</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In our second experiment (Tab. 3), we used 14 datasets, most of which are non-projective, from the CoNLL 2006 and 2008 shared tasks (Buchholz and Marsi, 2006; Surdeanu et al., 2008). Our third-order model achieved the best reported scores for English, Czech, German, and Dutch—which includes the three largest datasets and the ones with the most non-projective dependencies—and is on par with the state of the art for the remaining languages. To our knowledge, the speeds are the highest reported among higher-order non-projective parsers, and only about 3–4 times slower than the vine parser of Rush and Petrov (2012), which has lower accuracies.

5 Conclusions

We presented new third-order non-projective parsers which are both fast and accurate. We decoded with AD3, an accelerated dual decomposition algorithm which we adapted to handle large components, including specialized head automata for the third-order features, and a sequence model for head bigrams. Results are above the state of the art for large datasets and non-projective languages. In the hope that other researchers may find our implementation useful or are willing to contribute with further improvements, we made our parsers publically available as open source software.

Acknowledgments

We thank all reviewers for their insightful comments and Lingpeng Kong for help in converting the Penn Treebank to Stanford dependencies. This work was partially supported by the EU/FEDER programme, QREN/PORT Lisboa (Portugal), under the Intelligo project (contract 2012/24803), by a FCT grant PTDC/EEI-SII/2312/2012, and by NSF grant IIS-1054319.

References


