Minimal Dependency Length in Realization Ranking

Michael White and **Rajakrishnan Rajkumar**
Department of Linguistics
The Ohio State University
Columbus, OH, USA
{mwhite,raja}@ling.osu.edu

Abstract

Comprehension and corpus studies have found that the tendency to minimize dependency length has a strong influence on constituent ordering choices. In this paper, we investigate dependency length minimization in the context of discriminative realization ranking, focusing on its potential to eliminate egregious ordering errors as well as better match the distributional characteristics of sentence orderings in news text. We find that with a state-of-the-art, comprehensive realization ranking model, dependency length minimization yields statistically significant improvements in BLEU scores and significantly reduces the number of heavy/light ordering errors. Through distributional analyses, we also show that with simpler ranking models, dependency length minimization can go overboard, too often sacrificing canonical word order to shorten dependencies, while richer models manage to better counterbalance the dependency length minimization preference against (sometimes) competing canonical word order preferences.

1 Introduction

In this paper, we show that for the constituent ordering problem in surface realization, incorporating insights from the minimal dependency length theory of language production (Temperley, 2007) into a discriminative realization ranking model yields significant improvements upon a state-of-the-art baseline. We demonstrate empirically using OpenCCG, our CCG-based (Steedman, 2000) surface realization system, the utility of a global feature encoding the total dependency length of a given derivation. Although other works in the realization literature have used phrase length or head-dependent distances in their models (Filippova and Strube, 2009; Velldal and Oepen, 2005; White and Rajkumar, 2009, i.a.), to the best of our knowledge, this paper is the first to use insights from the minimal dependency length theory directly and study their effects, both qualitatively and quantitatively.

The impetus for this paper was the discovery that despite incorporating a sophisticated syntactic model borrowed from the parsing literature—including features with head-dependent distances at various scales—White & Rajkumar's (2009) realization ranking model still often performed poorly on weight-related decisions such as when to employ heavy-NP shift. Table 1 illustrates this point. In wsj_0034.9, the full model (incorporating numerous syntactic features) succeeds in reproducing the reference sentence, which is clearly preferable to the rather awkward variant selected by the baseline model (using various n-gram models). However, in wsj_0013.16, the full model fails to shift the temporal modifier *for now* next to the phrasal verb *turned down*, leaving it at the end of its very long verb phrase where it is highly ambiguous (with multiple intervening attachment sites). Conversely, in wsj_0044.3, the full model shifts *before* next to the verb, despite the NP *cheating* being very light, yielding a very confusing ordering given that *before* is meant to be intransitive.

The syntactic features in White & Rajkumar's (2009) realization ranking model are taken from Clark & Curran's (2007) normal form model

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 244–255, Jeju Island, Korea, 12–14 July 2012.

wsj_0034.9	they fell **into oblivion after the 1929 crash** .
FULL	[same]
BASELINE	they fell *after the 1929 crash into oblivion* .
wsj_0013.16	separately , the Federal Energy Regulatory Commission [$_{VP}$ turned down **for now** [$_{NP}$ a request by Northeast [$_{VP}$ seeking approval of [$_{NP}$ its possible purchase of PS of New Hampshire]]]] .
FULL	separately , the Federal Energy Regulatory Commission [$_{VP}$ turned down [$_{NP}$ a request by Northeast [$_{VP}$ seeking approval of [$_{NP}$ its possible purchase of PS of New Hampshire]]] *for now*] .
wsj_0044.3	she had seen **cheating before** , but these notes were uncanny .
FULL	she had seen *before cheating* , but these notes were uncanny .

Table 1: Examples of OpenCCG output with White & Rajkumar's (2009) models—the first represents a successful case, the latter two egregious ordering errors

(Table 3; see Section 3). In this model, head-dependenct distances are considered in conjunction with lexicalized and unlexicalized CCG derivation steps, thereby appearing in numerous features. As such, the model takes into account the interaction of dependency length with derivation steps, but in essence **does not consider the main effect of dependency length** itself. In this light, our investigation of dependency length minimization can be viewed as examining the question of whether realization ranking models can be made more accurate—and in particular, avoid egregious ordering errors—by incorporating a feature to account for the main effect of dependency length.

It is important to observe at this point that dependency length minimization is more of a preference than an optimization objective, which must be balanced against other order preferences at times. A closer reading of Temperley's (2007) study reveals that dependency length can sometimes run counter to many canonical word order choices. A case in point is the class of examples involving pre-modifying adjunct sequences that precede both the subject and the verb. Assuming that their parent head is the main verb of the sentence, a long-short sequence would minimize overall dependency length. However, in 613 examples found in the Penn Treebank, the average length of the first adjunct was 3.15 words while the second adjunct was 3.48 words long, thus reflecting a short-long pattern, as illustrated in the Temperley p.c. example in Table 2. Apart from these, Hawkins (2001) shows that arguments are generally located closer to the verb than adjuncts. Gildea and Temperley (2007) also suggest that adverb placement might involve cases which go against dependency length minimization. An examination of 295 legitimate long-short post-verbal constituent orders (counter to dependency length) from Section 00 of the Penn Treebank revealed that temporal adverb phrases are often involved in long-short orders, as shown in wsj_0075.13 in Table 2. In our setup, the preference to minimize dependency length can be balanced by features capturing preferences for alternate choices (e.g. the argument-adjunct distinction in our dependency ordering model, Table 4). Via distributional analyses, we show that while simpler realization ranking models can go overboard in minimizing dependency length, richer models largely succeed in overcoming this issue, while still taking advantage of dependency length minimization to avoid egregious ordering errors.

2 Background

2.1 Minimal Dependency Length

Comprehension and corpus studies (Gibson, 1998; Gibson, 2000; Temperley, 2007) point to the tendency of production and comprehension systems to adhere to principles of dependency length minimization. The idea of dependency length minimization is based on Gibson's (1998) Dependency Locality Theory (DLT) of comprehension, which predicts that longer dependencies are more difficult to process. DLT predictions have been further validated using comprehension studies involving eye-tracking corpora (Demberg and Keller, 2008). DLT metrics also correlate reasonably well with activation decay over time expressed in computational models of

Temperley (p.c.)	[In 1976], [as a film student at the Purchase campus of the State University of New York], Mr. Lane, shot ...
wsj_0075.13	The Treasury also said it plans to sell [$ 10 billion] [in 36-day cash management bills] [on Thursday].

Table 2: Counter-examples to dependency length minimization

comprehension (Lewis et al., 2006; Lewis and Vasishth, 2005).

Extending these ideas from comprehension, Temperley (2007) poses the question: Does language production reflect a preference for shorter dependencies as well so as to facilitate comprehension? By means of a study of Penn Treebank data, Temperley shows that English sentences do display a tendency to minimize the sum of all their head-dependent distances as illustrated by a variety of constructions. Further, Gildea and Temperley (2007) report that random linearizations have higher dependency lengths compared to actual English, while an "optimal" algorithm (from the perspective of dependency length minimization), which places dependents on either sides of a head in order of increasing length, is closer to actual English. Tily (2010) also applies insights from the above cited papers to show that dependency length constitutes a significant pressure towards language change. For head-final languages (e.g., Japanese), dependency length minimization results in the "long-short" constituent ordering in language production (Yamashita and Chang, 2001). More generally, Hawkins's (1994; 2000) processing domains, dependency length minimization and end-weight effects in constituent ordering (Wasow and Arnold, 2003) are all very closely related. The dependency length hypothesis goes beyond the predictions made by Hawkins' *Minimize Domains* principle in the case of English clauses with three post-verbal adjuncts: Gibson's DLT correctly predicts that the first constituent tends to be shorter than the second, while Hawkins' approach does not make predictions about the relative orders of the first two constituents.

However, it would be very reductive to consider dependency length minimization as the sole factor in language production. In fact, a large body of prior work discusses a variety of other factors involved in language production. These other preferences are either correlated with dependency length or can override the minimal dependency length preference. Complexity (Wasow, 2002; Wasow and Arnold, 2003), animacy (Snider and Zaenen, 2006; Branigan et al., 2008), information status considerations (Wasow and Arnold, 2003; Arnold et al., 2000), the argument-adjunct distinction (Hawkins, 2001) and lexical bias (Wasow and Arnold, 2003; Bresnan et al., 2007) are a few prominent factors. More recently, Anttila et al. (2010) argued that the principle of end weight can be revised by calculating weight in prosodic terms to provide more explanatory power. As Temperley (2007) suggests, a satisactory model should combine insights from multiple approaches, a theme which we investigate in this work by means of a rich feature set adapted from the parsing and realization literature. Our feature design has been inspired by the conclusions of the above-cited works pertaining to the role of dependency length minimization in syntactic choice in conjuction with other factors influencing constituent order. However, going beyond Temperley's corpus study, we confirm the utility of incorporating a feature for minimizing dependency length into machine-learned models with hundreds of thousands of features found to be useful in previous parsing and realization work, and investigate the extent to which these features can counterbalance a dependency length minimization preference in cases where canonical word order considerations should prevail.

2.2 Surface Realization with Combinatory Categorial Grammar (CCG)

We provide here a brief overview of CCG and the OpenCCG realizer; for further details, see the works cited below.

CCG (Steedman, 2000) is a unification-based categorial grammar formalism defined almost entirely in terms of lexical entries that encode sub-

Feature Type	Example
LexCat + Word	$s/s/np$ + before
LexCat + POS	$s/s/np$ + IN
Rule	$s_{dcl} \rightarrow np\ s_{dcl}\backslash np$
Rule + Word	$s_{dcl} \rightarrow np\ s_{dcl}\backslash np$ + bought
Rule + POS	$s_{dcl} \rightarrow np\ s_{dcl}\backslash np$ + VBD
Word-Word	$\langle$company, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$, bought$\rangle$
Word-POS	$\langle$company, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$, VBD$\rangle$
POS-Word	$\langle$NN, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$, bought$\rangle$
Word + Δ_w	$\langle$bought, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_w$
POS + Δ_w	$\langle$VBD, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_w$
Word + Δ_p	$\langle$bought, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_p$
POS + Δ_p	$\langle$VBD, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_p$
Word + Δ_v	$\langle$bought, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_v$
POS + Δ_v	$\langle$VBD, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_v$

Table 3: Basic and dependency features from Clark & Curran's (2007) normal form model; distances are in intervening words, punctuation marks and verbs, and are capped at 3, 3 and 2, respectively

categorization as well as syntactic features (e.g. number and agreement). OpenCCG is a parsing/generation library which includes a hybrid symbolic-statistical chart realizer (White, 2006; White and Rajkumar, 2009). The input to the OpenCCG realizer is a semantic graph, where each node has a lexical predication and a set of semantic features; nodes are connected via dependency relations. Internally, such graphs are represented using Hybrid Logic Dependency Semantics (HLDS), a dependency-based approach to representing linguistic meaning (Baldridge and Kruijff, 2002). Alternative realizations are ranked using integrated n-gram or averaged perceptron scoring models. In the experiments reported below, the inputs are derived from the gold standard derivations in the CCGbank (Hockenmaier and Steedman, 2007), and the outputs are the highest-scoring realizations found during the realizer's chart-based search.[1]

3 Feature Design

In the realm of paraphrasing using tree linearization, Kempen and Harbusch (2004) explore features which have later been appropriated into classification approaches for surface realization (Filippova and Strube, 2007). Prominent features include information status, animacy and phrase length. In the case of ranking models for surface realization, by far the most comprehensive experiments involving linguistically motivated features are reported in work of Cahill for German realization ranking (Cahill et al., 2007; Cahill and Riester, 2009). Apart from language model and Lexical Functional Grammar (LFG) c-structure and f-structure based features, Cahill also designed and incorporated features modeling information status considerations.

The feature sets explored in this paper extend those in previous work on realization ranking with OpenCCG using averaged perceptron models (White and Rajkumar, 2009; Rajkumar et al., 2009; Rajkumar and White, 2010) to include more comprehensive ordering features. The feature classes are listed below, where DEPLEN, HOCKENMAIER and DEPORD are novel, and the rest are as in earlier OpenCCG models. The inclusion of the DEPORD features is intended to yield a model with a similarly rich set of ordering features as Cahill and Forster's (2009) realization ranking model for German. Except where otherwise indicated, features are integer-valued, representing counts of occurrences in a derivation.

DEPLEN The total of the length between all semantic heads and dependents for a realization, where length is in intervening words[2] excluding punctuation. For length purposes, collapsed named entities were counted as a single word in the experiments reported here.

NGRAMS The log probabilities of the word sequence scored using three different n-gram models: a trigram word model, a trigram word model with named entity classes replacing words, and a trigram model over POS tags and supertags.

HOCKENMAIER As an extra component of the generative baseline, the log probability of the derivation according to (a reimplementation

[1] The realizer can also be run using inputs derived from OpenCCG's parser, though informal experiments suggest that parse errors tend to decrease generation quality.

[2] We also experimented with two other definitions of dependency length described in the literature, namely (1) counting only nouns and verbs to approximate counting by discourse referents (Gibson, 1998) and (2) omitting function words to approximate prosodic weight (Anttila et al., 2010); however, realization ranking accuracy was slightly worse than counting all non-punctuation words.

Feature Type	Example
HeadBroadPos + Rel + Precedes + HeadWord + DepWord	⟨VB, Arg0, dep, wants, he⟩
...+ HeadWord + DepPOS	⟨VB, Arg0, dep, wants, PRP⟩
...+ HeadPOS + DepWord	⟨VB, Arg0, dep, VBZ, he⟩
...+ HeadWord + DepPOS	⟨VB, Arg0, dep, VBZ, PRP⟩
HeadBroadPos + Side + DepWord1 + DepWord2	⟨NN, left, an, important⟩
...+ DepWord1 + DepPOS2	⟨NN, left, an, JJ⟩
...+ DepPOS1 + DepWord2	⟨NN, left, DT, important⟩
...+ DepPOS1 + DepPOS2	⟨NN, left, DT, JJ⟩
...+ Rel1 + Rel2	⟨NN, left, Det, Mod⟩

Table 4: Basic head-dependent and sibling dependent ordering features

of) Hockenmaier's (2003) generative syntactic model.

DISCRIMINATIVE NGRAMS Sequences from each of the n-gram models in the perceptron model.

AGREEMENT Features for subject-verb and animacy agreement as well as balanced punctuation.

C&C NF BASE The features from Clark & Curran's (2007) normal form model, listed in Table 3, minus the distance features.

C&C NF DISTANCE The distance features from the C&C normal form model, where the distance between a head and its dependent is measured in intervening words, punctuation marks or verbs; caps of 3, 3 and 2 (resp.) on the distances have the effect of binning longer distances.

DEPORD Several classes of features for ordering heads and dependents as well as sibling dependents on the same side of the head. The basic features—using words, POS tags and dependency relations, grouped by the broad POS tag of the head—are shown in Table 4. There are also similar features using words and a word class (instead of words and POS tags), where the class is either the named entity class, COLOR for color words, PRO for pronouns, one of 60-odd suffixes culled from the web, or HYPHEN or CAP for hyphenated or capitalized words. Additionally, there are features for detecting definiteness of an NP or PP (where the definiteness value is used in place of the POS tag).

Model	# Alph Feats	# Model Feats
GLOBAL	4	4
DEPLEN-GLOBAL	5	5
DEPORD-NONF	790,887	269,249
DEPORD-NODIST	1,035,915	365,287
DEPLEN-NODIST	1,035,916	366,094
DEPORD-NF	1,173,815	431,226
DEPLEN	1,173,816	428,775

Table 6: Model sizes—number of features in alphabet for each model (satisfying count cutoff of 5) along with number active in model after 5 training epochs

4 Evaluation

4.1 Experimental Conditions

We followed the averaged perceptron training procedure of White and Rajkumar (2009) with a couple of updates. First, as noted earlier, we used a reimplementation of Hockenmaier's (2003) generative syntactic model as an extra component of our generative baseline; and second, only five epochs of training were used, which was found to work as well as using additional epochs on the development set. As in the earlier work, the models were trained on the standard training sections (02–21) of an enhanced version of the CCGbank, using a lexico-grammar extracted from these sections.

The models tested in the experiments reported below are summarized in Table 5. The three groups of models are designed to test the impact of the dependency length feature when added to feature sets of increasing complexity. In more detail, the GLOBAL and DEPLEN-GLOBAL models contain dense features on entire derivations; their values are the log probabilities of the three n-gram mod-

Model	Dep Len	Ngram Mods	Hocken-maier	Discr Ngrams	Agree-ment	C&C NF Base	C&C NF Dist	Dep Ord
GLOBAL	N	Y	Y	N	N	N	N	N
DEPLEN-GLOBAL	Y	Y	Y	N	N	N	N	N
DEPORD-NONF	N	Y	Y	Y	Y	N	N	Y
DEPORD-NODIST	N	Y	Y	Y	Y	Y	N	Y
DEPLEN-NODIST	Y	Y	Y	Y	Y	Y	N	Y
DEPORD-NF	N	Y	Y	Y	Y	Y	Y	Y
DEPLEN	Y	Y	Y	Y	Y	Y	Y	Y

Table 5: Legend for experimental conditions

els used in the earlier work along with the Hockenmaier model (and the dependency length feature, in DEPLEN-GLOBAL). The second group is centered on DEPORD-NODIST, which contains all features except the dependency length feature and the distance features in Clark & Curran's normal form model, which may indirectly capture some dependency length minimization preferences. In addition to DEPLEN-NODIST—where the dependency length feature is added—this group also contains DEPORD-NONF, which is designed to test (as a side comparison) whether the Clark & Curran normal form base features are still useful even when used in conjunction with the new dependency ordering features. In the final group, DEPORD-NF contains all the features examined in this paper except the dependency length feature, while DEPLEN contains all the features including the dependency length feature. Note that the learned weight of the total dependency length feature was negative in each case, as expected.

Table 6 shows the sizes of the various models. For each model, the alphabet—whose size increases to over a million features—is the set of applicable features found to have discriminative value in at least 5 training examples; from these, a subset are made active (i.e., take on a non-zero weight) through perceptron updates when the feature value differs between the model-best and oracle-best realization.

4.2 BLEU Results

Following the usual practice in the realization ranking, we first evaluate our results quantitatively using exact matches and BLEU (Papineni et al., 2002), a corpus similarity metric developed for MT evaluation. Realization results for the development and test sections appear in Table 7. For all three model groups, the dependency length feature yields significant increases in BLEU scores, even in comparison to the model (DEPORD-NF) containing Clark & Curran's distance features in addition to the new dependency ordering features (as well as all other features but total dependency length). The second group additionally shows that the Clark & Curran normal form base features do indeed have a significant impact on BLEU scores even when used with

Model	% Exact	BLEU	Signif
Sect 00			
GLOBAL	33.03	0.8292	-
DEPLEN-GLOBAL	34.73	**0.8345**	***
DEPORD-NONF	42.33	0.8534	**
DEPORD-NODIST	43.12	0.8560	-
DEPLEN-NODIST	43.87	**0.8587**	***
DEPORD-NF	43.44	0.8590	-
DEPLEN	44.56	**0.8610**	**
Sect 23			
GLOBAL	34.75	0.8302	-
DEPLEN-GLOBAL	34.70	**0.8330**	***
DEPORD-NODIST	41.42	0.8561	-
DEPLEN-NODIST	42.95	**0.8603**	***
DEPORD-NF	41.32	0.8577	-
DEPLEN	42.05	**0.8596**	**

Table 7: Development (Section 00) & test (Section 23) set results—exact match percentage and BLEU scores, along with statistical significance of BLEU compared to the unmarked model in each group (* = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$); significant within-group winners (at $p < 0.05$) are shown in bold

Model	% DL Lower	% DL Greater	DL Mean	Signif
GOLD	n.a.	n.a.	41.02	-
GLOBAL	17.23	21.59	42.40	***
DEPLEN-GLOBAL	24.37	12.81	40.29	***
DEPORD-NONF	15.76	19.34	42.34	***
DEPORD-NODIST	14.58	19.06	42.03	***
DEPLEN-NODIST	17.75	14.82	40.87	n.s.
DEPORD-NF	14.96	17.65	41.58	***
DEPLEN	16.28	14.78	40.97	n.s.

Table 8: Dependency length compared to corpus—percentage of realizations with dependency length less than and greater than gold standard, along with mean dependency length, whose significance is tested against gold; 1671 development set (Section 00) complete realizations analyzed

the new dependency ordering model, as DEPORD-NONF is significantly worse than DEPORD-NODIST (the impact of the distance features is evident in the increases from the second group to the third group). As with the dev set, the dependency length feature yielded a significant increase in BLEU scores for each comparison on the test set also.

For each group, the statistical significance of the difference in BLEU scores between a model and the unmarked model (-) is determined by bootstrap resampling (Koehn, 2004).[3] Note that although the differences in BLEU scores are small, they end up being statistically significant because the models frequently yield the same top scoring realization, and reliably deliver improvements in the cases where they differ. In particular, note that DEPLEN and DEPORD-NF agree on the best realization 81% of the time, while DEPLEN-NODIST and DEPORD-NODIST have 78.1% agreement, and DEPLEN-GLOBAL and GLOBAL show 77.4% agreement; by comparison, DEPORD-NODIST and GLOBAL only agree on the best realization 51.1% of the time.

4.3 Detailed Analyses

The effect of the dependency length feature on the distribution of dependency lengths is illustrated in Table 8. The table shows the mean of the total dependency length of each realized derivation compared to the corresponding gold standard derivation, as well as the number of derivations with greater and lower dependency length. According to paired t-tests, the mean dependency lengths for the DEPLEN-NODIST and DEPLEN models do not differ significantly from the gold standard. In contrast, the mean dependency length of all the models that do not include the dependency length feature does differ significantly ($p < 0.001$) from the gold standard. Additionally, all these models have more realizations with dependency length greater than the gold standard, in comparison to the dependency length minimizing models; this shows the efficacy of the dependency length feature in approximating the gold standard. Interestingly, the DEPLEN-GLOBAL model significantly undershoots the gold standard on mean dependency length, and has the most skewed distribution of sentences with greater vs. lesser dependency length than the gold standard.

Model	% Short / Long	% Long / Short	% Eq	% Single Constit
GOLD	25.25	4.87	4.08	65.79
GLOBAL	23.15	7.86	3.94	65.04
DEPLEN-GLOBAL	24.58	5.57	4.09	65.76
DEPORD-NONF	23.13	6.61	4.03	66.23
DEPORD-NODIST	23.38	6.52	3.94	66.15
DEPLEN-NODIST	24.03	5.38	4.01	66.58
DEPORD-NF	23.74	5.92	3.96	66.40
DEPLEN	24.36	5.36	4.07	66.21

Table 9: Distribution of various kinds of post-verbal constituents in the development set (Section 00); 4692 gold cases considered

Apart from studying dependency length directly, we also looked at one of the attested effects of dependency length minimization, viz. the tendency to prefer short-long post-verbal constituents in production (Temperley, 2007). The relative lengths of adjacent post-verbal constituents were computed and their distribution is shown in Table 9. While calculating length, punctuation marks were excluded. Four kinds of constituents were found in the post-verbal domain. For every verb, apart from single constituents and equal length constituents, short-long and long-short sequences were also observed. Table 9 demonstrates that for both the gold standard corpus as well as the realizer models, short-long constituents were more frequent than long-short or equal length constituents. This follows the trend re-

[3] Kudos to Kevin Gimpel for making his resampling scripts available from `http://www.ark.cs.cmu.edu/MT/paired_bootstrap_v13a.tar.gz`.

Model	% Light / Heavy	% Heavy / Light	Signif
GOLD	8.60	0.36	-
GLOBAL	7.73	2.02	***
DEPLEN-GLOBAL	8.35	0.75	**
DEPORD-NONF	7.98	1.15	***
DEPORD-NODIST	8.04	1.12	***
DEPLEN-NODIST	8.23	0.45	n.s.
DEPORD-NF	8.26	0.71	**
DEPLEN	8.36	0.51	n.s.

Table 10: Distribution of heavy unequal constituents (length difference > 5) in Section 00; 4692 gold cases considered and significance tested against the gold standard using a χ-square test

ported by previous corpus studies of English (Temperley, 2007; Wasow and Arnold, 2003). The figures reported here show the tendency of the DEPLEN* models to be closer to the gold standard than the other models, especially in the case of short-long constituents.

We also performed an analysis of relative constituent lengths focusing on light-heavy and heavy-light cases; specifically, we examined unequal length constituent sequences where the length difference of the constituents was greater than 5, and the shorter constituent was under 5 words. Table 10 shows the results. Using a χ-square test, the distribution of heavy unequal length constituent counts in the DEPLEN-NODIST and DEPLEN models does not significantly differ from that of the gold standard. In contrast, for all the other models, the counts do differ significantly from the gold standard.

4.4 Examples

Table 11 shows examples of how the dependency length feature (DEPLEN) affects the output even in comparison to a model (DEPORD) with a rich set of discriminative syntactic and dependency ordering features, but no features directly targeting relative weight. In wsj_0015.7, the dependency length model produces an exact match, while the DEPORD model fails to shift the short temporal adverbial *next year* next to the verb, leaving a confusingly repetitive *this year next year* at the end of the sentence. In wsj_0020.1, the dependency length model produces a nearly exact match with just an equally acceptable inversion of *closely watching*. By contrast, the DEPORD model mistakenly shifts the direct object *South Korea, Taiwan and Saudia Arabia* to the end of the sentence where it is difficult to understand following two very long intervening phrases. In wsj_0021.8, both models mysteriously put *not* in front of the auxiliary and leave out the complementizer, but DEPORD also mistakenly leaves *before* at the end of the verb phrase where it is again apt to be interpreted as modifying the preceding verb. In wsj_0075.13, both models put the temporal modifier *on Thursday* in its canonical VP-final position, despite this order running counter to dependency length minimization. Finally, wsj_0014.2 shows a case where DEPORD is nearly an exact match (except for a missing comma), but the dependency length model fronts the PP *on the 12-member board*, where it is grammatical but rather marked (and not motivated in the discourse context).

4.5 Interim Discussion

The experiments show a consistent positive effect of the dependency length feature in improving BLEU scores and achieving a better match with the corpus distributions of dependency length and short/long constituent orders. The results in Table 10 are particulary encouraging, as they show that minimizing dependency length reduces the number of realizations in which a heavy constituent precedes a light one down to essentially the level of the corpus, thereby eliminating many realizations that can be expected to have egregious errors like those shown in Table 11.

Intriguingly, there is some evidence that a negatively weighted total dependency length feature can go too far in minimizing dependency length, in the absence of other informative features to counterbalance it. In particular, the DEPLEN-GLOBAL model in Table 8 has significantly lower dependency length than the corpus, but in the richer models with discriminative synactic and dependency ordering features, there are no significant differences. It may still be though that additional features are necessary to counteract the tendency towards dependency length minimization, for example to ensure that initial constituents play their intended role in establishing and continuing topics in discourse, as also observed in Table 11.

wsj_0015.7	the exact amount of the refund will be determined **next year** based on actual collections made until Dec. 31 of this year .
DEPLEN	[same]
DEPORD	the exact amount of the refund will be determined based on actual collections made until Dec. 31 of this year *next year* .
wsj_0020.1	the U.S. , claiming some success in its trade diplomacy , removed South Korea , Taiwan and Saudi Arabia from a list of countries it is closely watching for allegedly failing to honor U.S. patents , copyrights and other intellectual-property rights .
DEPLEN	the U.S. claiming some success in its trade diplomacy , removed **South Korea , Taiwan and Saudi Arabia** from a list of countries it is *watching closely* for allegedly failing to honor U.S. patents , copyrights and other intellectual-property rights .
DEPORD	the U.S. removed from a list of countries it is *watching closely* for allegedly failing to honor U.S. patents , copyrights and other intellectual-property rights , claiming some success in its trade diplomacy , *South Korea , Taiwan and Saudi Arabia* .
wsj_0021.8	but he has not said before that the country wants half the debt forgiven .
DEPLEN	but he *not* has said **before** ∅ the country wants half the debt forgiven .
DEPORD	but he *not* has said ∅ the country wants half the debt forgiven **before** .
wsj_0075.13	The Treasury also said it plans to sell [$ 10 billion] [in 36-day cash management bills] [on Thursday].
DEPLEN	[same]
DEPORD	[same]
wsj_0014.2	they succeed Daniel M. Rexinger , retired Circuit City executive vice president , and Robert R. Glauber , U.S. Treasury undersecretary , on the 12-member board .
DEPORD	they succeed Daniel M. Rexinger , retired Circuit City executive vice president , and Robert R. Glauber , U.S. Treasury undersecretary ∅ on the 12-member board .
DEPLEN	*on the 12-member board* they succeed Daniel M. Rexinger , retired Circuit City executive vice president , and Robert R. Glauber , U.S. Treasury undersecretary .

Table 11: Examples of realized output for full models with and without the dependency length feature

4.6 Targeted Human Evaluation

To determine whether heavy-light ordering differences often represent ordering errors (including egregious ones), rather than simply representing acceptable variation, we conducted a targeted human evaluation on examples of this kind. Specifically, for each of the DEPLEN* models and their corresponding models without the dependency length feature, we chose the 25 sentences from the development section whose realizations exhibited the greatest difference in dependency length between sibling constituents appearing in opposite orders, and asked two judges (not the authors) to choose which of the two realizations best expressed the meaning of the reference sentence in a grammatical and fluent way, with the choice forced (2AFC). Table 12 shows the results. Agreement between the judges was high,

Model	% Preferred	% Agr	Signif
GLOBAL	22	-	-
DEPLEN-GLOBAL	78	84	***
DEPORD-NODIST	24	-	-
DEPLEN-NODIST	76	92	***
DEPORD-NF	26	-	-
DEPLEN	74	96	***

Table 12: Targeted human evaluation—percentage of realizations preferred by two human judges in a 2AFC test among the 25 development set sentences with the greatest differences in dependency length, with a binomial test for significance

with only one disagreement on the realizations from the DEPLEN and DEPORD-NF models (involving an acceptable paraphrase in our judgment), and only four disagreements on the DEPLEN-GLOBAL and GLOBAL realizations. Pooling the judgments, the preference for the DEPLEN* models was well above the chance level of 50% according to a binomial test ($p < 0.001$ in each case). Inspecting the data ourselves, we found that many of the items did indeed involve egregious ordering errors that the DEPLEN* models managed to avoid.

5 Related Work

As noted in the introduction, to the best of our knowledge this paper is the first to examine the impact of dependency length minimization on realization ranking. While there have been quite a few papers to date reporting results on Penn Treebank data, since the various systems make different assumptions regarding the specificity of their inputs, all but the most broad-brushed comparisons remain impossible at present, and thus detailed studies such as the present one can only be made within the context of different models for the same system. Some progress on this issue has been made in the context of the Generation Challenges Surface Realization Shared Task (Belz et al., 2011), but it remains to be seen to what extent fair cross-system comparisons using common inputs can be achieved.

For (very) rough comparison purposes, Table 13 lists our results in the context of those reported for various other systems on PTB Section 23. As the table shows, the OpenCCG scores are quite competitive, exceeded only by Callaway's (2005) extensively hand-crafted system as well as Bohnet et al.'s (2011) system on shared task shallow inputs (-S), which performs much better than their system on deep inputs (-D) that more closely resemble OpenCCG's.

System	Coverage	BLEU	% Exact
Callaway (05)	98.5%	0.9321	57.5
Bohnet et al.-S (11)	100%	0.8911	-
OpenCCG (12)	97.1%	0.8596	42.1
OpenCCG (09)	97.1%	0.8506	40.5
Ringger et al. (04)	100%	0.836	35.7
Bohnet et al.-D (11)	100%	0.7943	-
Langkilde-Geary (02)	83%	0.757	28.2
Guo et al. (08)	100%	0.7440	19.8
Hogan et al. (07)	≈100%	0.6882	-
OpenCCG (08)	96.0%	0.6701	16.0
Nakanishi et al. (05)	90.8%	0.7733	-

Table 13: PTB Section 23 BLEU scores and exact match percentages in the NLG literature (Nakanishi et al.'s results are for sentences of length 20 or less)

6 Conclusions

In this paper, we have investigated dependency length minimization in the context of realization ranking, focusing on its potential to eliminate egregious ordering errors as well as better match the distributional characteristics of sentence orderings in news text. When added to a state-of-the-art, comprehensive realization ranking model, we showed that including a dense, global feature for minimizing total dependency length yields statistically significant improvements in BLEU scores and significantly reduces the number of heavy-light ordering errors. Going beyond the BLEU metric, we also conducted a targeted human evaluation to confirm the utility of the dependency length feature in models of varying richness. Interestingly, even with the richest model, in some cases we found that the dependency length feature still appears to go too far in minimizing dependency length, suggesting that further counter-balancing features—especially ones for the sentence-initial position (Filippova and Strube, 2009)—warrant investigation in future work.

Acknowledgments

This work was supported in part by NSF grants no. IIS-1143635 and IIS-0812297. We thank the anonymous reviewers for helpful comments and discussion, and Scott Martin and Dennis Mehay for their participation in the targeted human evaluation.

References

Arto Anttila, Matthew Adams, and Mike Speriosu. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes*.

Jennifer E. Arnold, Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76:28–55.

Jason Baldridge and Geert-Jan Kruijff. 2002. Coupling CCG and Hybrid Logic Dependency Semantics. In *Proc. ACL-02*.

Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 217–226, Nancy, France, September. Association for Computational Linguistics.

Bernd Bohnet, Simon Mille, Benoît Favre, and Leo Wanner. 2011. <stumaba >: From deep representation to surface. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 232–235, Nancy, France, September. Association for Computational Linguistics.

H Branigan, M Pickering, and M Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118(2):172–189.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the Dative Alternation. *Cognitive Foundations of Interpretation*, pages 69–94.

Aoife Cahill and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of*, ACL-IJCNLP '09, pages 817–825, Morristown, NJ, USA. Association for Computational Linguistics.

Aoife Cahill, Martin Forst, and Christian Rohrer. 2007. Designing features for parse disambiguation and realisation ranking. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the 12th International Lexical Functional Grammar Conference*, pages 128–147. CSLI Publications, Stanford.

Charles Callaway. 2005. The types and distributions of errors in a wide coverage surface realizer evaluation. In *Proceedings of the 10th European Workshop on Natural Language Generation*.

Stephen Clark and James R. Curran. 2007. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. *Computational Linguistics*, 33(4):493–552.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Katja Filippova and Michael Strube. 2007. Generating constituent order in German clauses. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computer Linguistics.

Katja Filippova and Michael Strube. 2009. Tree linearization in English: Improving language model based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 225–228, Boulder, Colorado, June. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Edward Gibson. 2000. Dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*. MIT Press, Cambridge, MA.

Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic, June. Association for Computational Linguistics.

Yuqing Guo, Josef van Genabith, and Haifeng Wang. 2008. Dependency-based n-gram models for general purpose sentence realisation. In *Proc. COLING-08*.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press, New York.

John A. Hawkins. 2000. The relative order of prepositional phrases in English: Going beyond manner-place-time. *Language Variation and Change*, 11(03):231–266.

John A. Hawkins. 2001. Why are categories adjacent? *Journal of Linguistics*, 37:1–34.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Julia Hockenmaier. 2003. *Data and models for statistical parsing with Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Deirdre Hogan, Conor Cafferkey, Aoife Cahill, and Josef van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. In *Proc. EMNLP-CoNLL*.

Gerard Kempen and Karin Harbusch. 2004. Generating natural word orders in a semi-free word order language: Treebank-based linearization preferences for German. In Alexander F. Gelbukh, editor, *CICLing*, volume 2945 of *Lecture Notes in Computer Science*, pages 350–354. Springer.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Irene Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proc. INLG-02*.

R. L. Lewis and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:1–45, May.

Richard L. Lewis, Shravan Vasishth, and Julie Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

Hiroko Nakanishi, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic methods for disambiguation of an HPSG-based chart generator. In *Proc. IWPT-05*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL-02*.

Rajakrishnan Rajkumar and Michael White. 2010. Designing agreement features for realization ranking. In *Coling 2010: Posters*, pages 1032–1040, Beijing, China, August. Coling 2010 Organizing Committee.

Rajakrishnan Rajkumar, Michael White, and Dominic Espinosa. 2009. Exploiting named entity classes in CCG surface realization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 161–164, Boulder, Colorado, June. Association for Computational Linguistics.

Eric Ringger, Michael Gamon, Robert C. Moore, David Rojas, Martine Smets, and Simon Corston-Oliver. 2004. Linguistically informed statistical models of constituent structure for ordering in sentence realization. In *Proc. COLING-04*.

Neal Snider and Annie Zaenen. 2006. Animacy and syntactic structure: Fronted NPs in English. In M. Butt, M. Dalrymple, and T.H. King, editors, *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*. CSLI Publications, Stanford.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300 – 333.

Harry Tily. 2010. *The Role of Processing Complexity in Word Order Variation and Change*. Ph.D. thesis, Stanford University.

Erik Velldal and Stefan Oepen. 2005. Maximum entropy models for realization ranking. In *Proc. MT-Summit X*.

Thomas Wasow and Jennifer Arnold. 2003. *Post-verbal Constituent Ordering in English*. Mouton.

Tom Wasow. 2002. *Postverbal Behavior*. CSLI Publications, Stanford.

Michael White and Rajakrishnan Rajkumar. 2009. Perceptron reranking for CCG realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Singapore, August. Association for Computational Linguistics.

Michael White. 2006. Efficient Realization of Coordinate Structures in Combinatory Categorial Grammar. *Research on Language & Computation*, 4(1):39–75.

Hiroko Yamashita and Franklin Chang. 2001. "Long before short" preference in the production of a head-final language. *Cognition*, 81.