Linguistically Motivated Complementizer Choice in Surface Realization

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

Abstract

This paper shows that using linguistically motivated features for English *that*-complementizer choice in an averaged perceptron model for classification can improve upon the prediction accuracy of a state-of-the-art realization ranking model. We report results on a binary classification task for predicting the presence/absence of a *that*-complementizer using features adapted from Jaeger's (2010) investigation of the uniform information density principle in the context of *that*-mentioning. Our experiments confirm the efficacy of the features based on Jaeger's work, including information density–based features. The experiments also show that the improvements in prediction accuracy apply to cases in which the presence of a *that*-complementizer arguably makes a substantial difference to fluency or intelligiblity. Our ultimate goal is to improve the performance of a ranking model for surface realization, and to this end we conclude with a discussion of how we plan to combine the local complementizer-choice features with those in the global ranking model.

1 Introduction

Johnson (2009) observes that in developing statistical parsing models, "shotgun" features — that is, myriad scattershot features that pay attention to superficial aspects of structure — tend to be remarkably useful, while features based on linguistic theory seem to be of more questionable utility, with the most basic linguistic insights tending to have the greatest impact.[1] Johnson also notes that feature design is perhaps the most important but least understood aspect of statistical parsing, and thus the disappointing impact of linguistic theory on parsing models is of real consequence. In this paper, by contrast, we show that in the context of surface realization, using linguistically motivated features for English *that*-complementizer choice can improve upon the prediction accuracy of a state-of-the-art realization ranking model, arguably in ways that make a substantial difference to fluency and intelligiblity.[2] In particular, we report results on a binary classification task for predicting the presence or absence of a *that*-complementizer using features adapted from Jaeger's (2010) investigation of the **uniform information density** principle in the context of *that*-mentioning. This information-theoretic principle predicts that language production is affected by a preference to distribute information uniformly across the linguistic signal. In Jaeger's study, uniform information density emerges as an important predictor of speakers' syntactic reduction preferences even when taking a sizeable variety of controls based on competing hypotheses into account. Our experiments confirm the efficacy of the features based on Jaeger's work, including information density–based features.

[1] The term *"shotgun" feature* appears in the slides for Johnson's talk (http://www.cog.brown.edu/~mj/papers/johnson-eacl09-workshop.pdf), rather than in the paper itself.

[2] For German surface realization, Cahill and Riester (2009) show that incorporating information status features based on the linguistics literature improves performance on realization ranking.

Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop, pages 39–44,
Edinburgh, Scotland, UK, July 31, 2011.

That-complementizers are optional words that introduce sentential complements in English. In the Penn Treebank, they are left out roughly two-thirds of the time, thereby enhancing conciseness. This follows the low complementizer rates reported in previous work (Tagliamonte and Smith, 2005; Cacoullos and Walker, 2009). While some surface realizers, such as FUF/SURGE (Elhadad, 1991), have made use of input features to control the choice of whether to include a *that*-complementizer, for many applications the decision seems best left to the realizer, since multiple surface syntactic factors appear to govern the choice, rather than semantic ones. In our experiments, we use the OpenCCG[3] surface realizer with logical form inputs underspecified for the presence of *that* in complement clauses. While in many cases, adding or removing *that* results in an acceptable paraphrase, in the following example, the absence of *that* in (2) introduces a local ambiguity, which the original Penn Treebank sentence avoids by including the complementizer.

(1) He said that for the second month in a row, food processors reported a shortage of nonfat dry milk. (WSJ0036.61)

(2) ? He said for the second month in a row, food processors reported a shortage of nonfat dry milk.

The starting point for this paper is White and Rajkumar's (2009) realization ranking model, a state-of-the-art model employing shotgun features galore. An error analysis of this model, performed by comparing CCGbank Section 00 realized derivations with their corresponding gold standard derivations, revealed that out of a total of 543 *that*-complementizer cases, the realized output did not match the gold standard choice 82 times (see Table 3 in Section 5 for details). Most of these mismatches involved cases where a clause originally containing a *that*-complementizer was realized in reduced form, with no *that*. This under-prediction of *that*-inclusion is not surprising, since the realization ranking model makes use of baseline n-gram model features, and n-gram models are known to have a built-in bias for strings with fewer words.

[3] openccg.sf.net

We report here on experiments comparing this global model to ones that employ local features specifically designed for *that*-choice in complement clauses. As a prelude to incorporating these features into a model for realization ranking, we study the efficacy of these features in isolation by means of a binary classification task to predict the presence/absence of *that* in complement clauses. In a global realization ranking setting, the impact of these phenomenon-specific features might be less evident, as they would interact with other features for lexical selection and ordering choices that the ranker makes. Note that a comprehensive ranking model is desirable, since linear ordering and *that*-complementizer choices may interact. For example, Hawkins (2003) reports examples where explicitly marked phrases can occur either close to or far from their heads as in (3) and (4), whereas zero-marked phrases are only rarely attested at some distance from their heads and prefer adjacency, as (5) and (6) show.

(3) I realized [that he had done it] with sadness in my heart.

(4) I realized with sadness in my heart [that he had done it].

(5) I realized [he had done it] with sadness in my heart.

(6) ? I realized with sadness in my heart [he had done it].

2 Background

CCG (Steedman, 2000) is a unification-based categorial grammar formalism defined almost entirely in terms of lexical entries that encode subcategorization 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). The chart realizer takes as input logical forms represented internally using Hybrid Logic Dependency Semantics (HLDS), a dependency-based approach to representing linguistic meaning (Baldridge and Kruijff, 2002). To illustrate the input to OpenCCG, consider the semantic dependency graph in Figure 1. In the graph, each node has a lexical predication (e.g. **make.03**) and a set of semantic features (e.g.

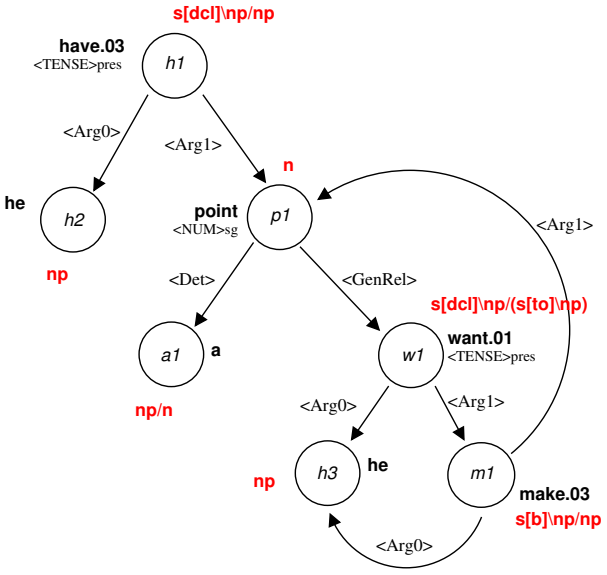

Figure 1: Semantic dependency graph from the CCGbank for *He has a point he wants to make [. . .]*, along with gold-standard supertags (category labels)

⟨NUM⟩sg); nodes are connected via dependency relations (e.g. ⟨ARG0⟩). In HLDS, each semantic head (corresponding to a node in the graph) is associated with a nominal that identifies its discourse referent, and relations between heads and their dependents are modeled as modal relations. We extract HLDS-based quasi logical form graphs from the CCGbank and semantically empty function words such as complementizers, infinitival-*to*, expletive subjects, and case-marking prepositions are adjusted to reflect their purely syntactic status. Alternative realizations are ranked using an averaged perceptron model described in the next section.

3 Feature Design

White and Rajkumar's (2009) realization ranking model serves as the baseline for this paper. It is a global, averaged perceptron ranking model using three kinds of features: (1) the log probability of the candidate realization's word sequence according to three linearly interpolated language models (as well as a feature for each component model), much as in the log-linear models of Velldal & Oepen (2005) and Nakanishi et al. (2005); (2) integer-valued syntactic features, representing counts of occurrences in a derivation, from Clark & Curran's (2007) normal form model; and (3) discriminative n-gram features (Roark et al., 2004), which count the occurrences of each n-gram in the word sequence.

Table 1 shows the new complementizer-choice features investigated in this paper. The example features mentioned in the table are taken from the two complement clause (CC) forms (with-*that* CC vs. *that*-less CC) of the sentence below:

(7) The finding probably will support those who **argue** [*that*/∅ the U.S. should regulate the class of asbestos including crocidolite more stringently than the common kind of asbestos, chrysotile, found in most schools and other buildings], Dr. Talcott said. (WSJ0003.19)

The first class of features, dependency length and position of CC, have been adapted from the related control features in Jaeger's (2010) study. For the above example, the position of the matrix verb with respect to the start of the sentence (feature name *mvInd* and having the value 7.0), the distance between the matrix verb and the onset of the CC (feature name *mvCCDist* with the value 1.0) and finally the length of the CC (feature *ccLen* with value of 29.0 for the *that*-CC and 28.0 for the *that*-less CC) are encoded as features. The second class of features includes various properties of the matrix verb viz. POS tag, form, stem and supertag (feature names *mv Pos, mvStem, mvForm, mvSt*, respectively). These features were motivated by the fact that Jaeger controls for the per-verb bias of this construction, as attested in the earlier literature. The third class of features are related to information density. Jaeger (2010) estimates information density at the CC onset by using matrix verb subcategorization frequency. In our case, more like the n-gram features employed by Levy and Jaeger (2007), we used log probabilities from two existing n-gram models, viz. a trigram word model and trigram word model with semantic class replacement. For each CC, two features (one per language model) were extracted by calculating the average of the log probs of individual words from the beginning of the complement clause. In the *that*-CC version of the example above, local CC-features having the prefix *$uidCCMean* were calculated by averaging the individual log probs of the 3 words *that the U.S.* to get feature values of -0.8353556 and -2.0460036 per language model (see

Feature	Example for *that*-CCs	Example for *that*-less CCs
Dependency length and position of CC		
Position of matrix verb	thatCC:mvInd 7.0	noThatCC:mvInd 7.0
Dist between matrix verb & CC	thatCC:mvCCDist 1.0	noThatCC:mvCCDist 1.0
Length of CC	thatCC:ccLen 29.0	noThatCC:ccLen 28.0
Matrix verb features		
POS-tag	thatCC:mvPos:VBP 1.0	noThatCC:mvPos:VBP 1.0
Stem	thatCC:mvStem:argue 1.0	noThatCC:mvStem:argue 1.0
Form	thatCC:mvForm:argue 1.0	noThatCC:mvForm:argue 1.0
CCG supertag	thatCC:mvSt:s[dcl]\np/s[em] 1.0	noThatCC:mvSt:s[dcl]\np/s[dcl] 1.0
uniform information density (UID)		
Average n-gram log probs	thatCC:$uidCCMean1 -0.8353556	noThatCC:$uidCCMean1 -2.5177214
of first 2 words of *that-less* CCs	thatCC:$uidCCMean2 -2.0460036	noThatCC:$uidCCMean2 -3.6464245
or first 3 words of *that*-CCs		

Table 1: New features introduced (the prefix of each feature encodes the type of CC; subsequent parts supply the feature name)

last part of Table 1). In the *that*-less CC version, *$uidCCMean* features were calculated by averaging the log probs of the first two words in the complement clause, i.e. *the U.S.*

4 Classification Experiment

To train a local classification model to predict the presence of *that* in complement clauses, we used an averaged perceptron ranking model with the complementizer-specific features listed in Table 1 to rank alternate with-*that* vs. *that*-less CC choices. For each CC classification instance in CCGbank Sections 02–21, the derivation of the competing alternate choice was created; i.e., in the case of a *that*-CC, the corresponding *that*-less CC was created and vice versa. Table 2 illustrates classification results on Sections 00 (development) using models containing different feature sets & Section 23 (final test) for the best-performing classification and ranking models. For both the development as well as test sections, the local classification model performed significantly better than the global realization ranking model according to McNemar's χ^2 test ($p = 0.005$, two-tailed). Feature ablation tests on the development data (Section 00) revealed that removing the information density features resulted in a loss of accuracy of around 1.8%.

5 Discussion

As noted in the introduction, in many cases, adding or removing *that* to/from the corpus sentence results in an acceptable paraphrase, while in other cases the presence of *that* appears to make a substantial difference to intelligibility or fluency. In order to better understand the effect of the complementizer-specific features, we examined three construction types in the development data, viz. non-adjacent complement clauses, gerundive matrix verbs and a host of sub-cases involving a matrix *be*-verb (*wh*-clefts, *be*+adjective etc.), where the presence of *that* seemed to make the most difference. The results are provided in Table 3. As is evident, the global realization ranking model under-proposes the *that*-choice, most likely due to the preference of n-gram models towards fewer words, while the local classifica-

Model Features	% 00	% 23
Most Frequent Baseline	68.7	66.8
Global Realization Ranking	78.45	77.0
Local That-*Classification*		
Only UID feats	74.77	
Table 1 features except UID ones	81.4	
Both feature sets above	**83.24**	**83.02**

Table 2: Classification accuracy results (Section 00 has 170/543 *that*-CCs; Section 23 has 192/579 *that*-CCs)

Construction	*%that*	*% that* / %Accuracy	
	Gold	Classification	Ranking
Gerundive (26)	53.8	61.5 / 92.3	26.9 / 57.7
Be-verb (21)	71.4	95.2 / 66.7	47.6 / 57.1
Non-adjacent CCs (53)	49.1	54.7 / 67.9	30.2 / 66.0
Total (543)	31.3	29.3 / 83.2	21.9 / 78.5

Table 3: Section 00 construction-wise *that*-CC proportions and model accuracies (total CC counts given in brackets alongside labels); gold standard obviously has 100% accuracy; models are local *that*-classification and White and Rajkumar's (2009) global realization ranking model

WSJ0049.64	**Observing** [*that*/?∅ the judge has never exhibited any bias or prejudice], Mr. Murray concluded that he would be impartial in any case involving a homosexual or prostitute as a victim.
WSJ0020.16	" what this tells us **is** [*that*/?∅ U.S. trade law is working] ", he said .
WSJ0010.5	The idea, of course: to **prove** to 125 corporate decision makers [*that*/?∅ the buckle on the Rust Belt is n't so rusty after all , that it 's a good place for a company to expand].
WSJ0044.118	Editorials in the Greenville newspaper **allowed** [that/?∅ Mrs. Yeargin was wrong], but also said the case showed how testing was being overused.
WSJ0060.7	Viacom **denies** [∅/?*that* it 's using pressure tactics].
WSJ0018.4	The documents also **said** [*that*/?∅ although the 64-year-old Mr. Cray has been working on the project for more than six years , the Cray-3 machine is at least another year away from a fully operational prototype].

Table 4: Examples from model comparison

tion model is closer to the gold standard in terms of *that*-choice proportions. For all the three construction types as well as overall, classifier performance was better than ranker performance. The difference in performance between the local classification and global ranking models in the case of gerundive matrix verbs is statistically significant according to the McNemar's χ^2 test (Bonferroni corrected, two tailed $p = 0.001$). The performance difference was not significant with the other two constructions, however, using only the cases in Section 00.

Table 4 lists relevant examples where the classification model's *that*-choice prediction matched the gold standard while a competing model's prediction did not. Example WSJ0049.64 is one such instance of classifier success involving a gerundive matrix verb (in contrast to the realization ranking model), Example WSJ0020.16 exemplifies success with a *wh*-cleft construction and Example WSJ0010.5 contains a non-adjacent CC. Apart from these construction-based analyses, examples like WSJ0044.118 indicate that the classification model prefers the *that*-CC choice in cases that substantially improve intelligiblity, as here the overt complementizer helps to avoid a local syntactic ambiguity where the *NP* in *allowed NP* is unlikely to be interpreted as the start of an *S*.

Finally, we also studied the effect of the uniform information density features by comparing the full classification model to a model without the UID features. The full classification model exhibited a trend towards significantly outperforming the ablated model (McNemar's $p = 0.10$, 2-tailed); more test data would be needed to establish significance conclusively. Examples are shown at the bottom of Table 4. In WSJ0060.7, the full classification model predicted a *that*-less clause (matching the gold standard), while the ablated classification model predicted a clause with *that*. In all such examples except one, the information density features helped the classification model avoid predicting *that*-inclusion when not necessary. Example WSJ0018.4 is the only instance where the best classification model differed in predicting the *that*-choice.

6 Conclusions and Future Work

In this paper, we have shown that using linguistically motivated features for English *that*-complementizer choice in a local classifier can improve upon the prediction accuracy of a state-of-the-art global realization ranking model employing myriad shotgun features, confirming the efficacy of features based on Jaeger's (2010) investigation of the uniform information density principle in the context of *that*-mentioning. Since *that*-complementizer choice interacts with other realization decisions, in future work we plan to investigate incorporating these features into the global realization ranking model. This move will require binning the real-valued features, as multiple complement clauses can appear in a single sentence. Should feature-level integration prove ineffective, we also plan to investigate alternative architectures, such as using the local classifier outputs as features in the global model.

Acknowledgements

This work was supported in part by NSF IIS-0812297 and by an allocation of computing time from the Ohio Supercomputer Center. Our thanks also to Florian Jaeger, William Schuler, Peter Culicover and the anonymous reviewers for helpful comments and discussion.

References

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

Rena Torres Cacoullos and James A. Walker. 2009. On the persistence of grammar in discourse formulas: A variationist study of "that". *Linguistics*, 47(1):1–43.

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

Michael Elhadad. 1991. FUF: The universal unifier user manual version 5.0. Technical Report CUCS-038-91, Dept. of Computer Science, Columbia University.

John A. Hawkins. 2003. Why are zero-marked phrases close to their heads? In Günter Rohdenburg and Britta Mondorf, editors, *Determinants of Grammatical Variation in English*, Topics in English Linguistics 43. De Gruyter Mouton, Berlin.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage information density. *Cognitive Psychology*, 61(1):23–62, August.

Mark Johnson. 2009. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11, Athens, Greece, March. Association for Computational Linguistics.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, 19:849.

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

Brian Roark, Murat Saraclar, Michael Collins, and Mark Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proc. ACL-04*.

Mark Steedman. 2000. *The syntactic process*. MIT Press, Cambridge, MA, USA.

S. Tagliamonte and J. Smith. 2005. No momentary fancy! the zero 'complementizer' in English dialects. *English Language and Linguistics*, 9(2):289–309.

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

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 and Computation*, 4(1):39–75.