The Importance of Lexicalized Syntax Models for Natural Language Generation Tasks

Hal Daumé III, Kevin Knight, Irene Langkilde-Geary, Daniel Marcu and Kenji Yamada
Information Sciences Institute
Computer Science Department
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{hdaume,knight,ilangkil,marcu,kyamada}@isi.edu

Abstract

The parsing community has long recognized the importance of lexicalized models of syntax. By contrast, these models do not appear to have had an impact on the statistical NLG community. To prove their importance in NLG, we show that a lexicalized model of syntax improves the performance of a statistical text compression system, and show results that suggest it would also improve the performances of an MT application and a pure natural language generation system.

1 Introduction

We distinguish between three types of language models:

- n-gram language models look only at word sequences to gauge the quality of a sentence.
- Non-lexicalized syntax models consider only syntactic structure down to the level of word tags in assessing the grammaticality of a sentence. A PCFG is an example of such a model.
- Lexicalized syntax models take into account both sentence syntax and lexical values when determining the quality of a sentence.

The parsing community has long recognized the importance of lexicalized models of syntax for building robust natural language applications. For example, Charniak (1997) showed that by using a lexicalized syntax model instead of a non-lexicalized PCFG syntax model trained on Penn Treebank data, one can increase the performance of a syntactic parser from 73.75% labeled recall and precision to 83.75%. More sophisticated lexicalized models of syntax (Collins, 1997; Charniak, 2000) have increased the performance of syntactic parsers to 90% labeled recall and precision. Lexicalized models of syntax have been also proven useful in speech recognition (Chelba and Jelinek, 1998) and language modeling (Charniak, 2001; Roark, 2001).

By contrast, lexicalized models of syntax do not appear to have had an impact on the statistical NLG community. Langkilde and Knight (1998), for example, use an n-gram model to select between different lexical renderings of a meaning representation. Knight and Marcu (2000) use a combination of bigram and context free probabilities to select between sentence compressions. To our knowledge, the only NLG work that resonates with the work in parsing, speech recognition, and language modeling is that of Bangalore and Rambow (2000), who show that a statistical generation system that uses a lexicalized hierarchical model of syntax outperforms a system that uses a random model.

Given the small interest in exploiting lexicalized models of syntax in NLG, we may conclude that such models have no role to play in this area. In this paper, we show that this is not the case. To prove the importance of lexicalized models of syntax in NLG, we focus on three distinct tasks, each involving a generation component. First, we show that a lexicalized model of syntax can improve the performance of a statistics-based text compression system. Second, we show that a lexicalized model of syntax may improve the outputs of a Chinese-to-English machine translation system. Finally, we analyze the results of a pure natural language genera-

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2002		**00-00-2002 to 00-00-2002**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Importance of Lexicalized Syntax Models for Natural Language Generation Tasks	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Information Sciences Institute ,4676 Admiralty Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

tion system.[1]

In each experiment, we assess the impact that a lexicalized model of syntax may have on improving the quality of existing systems using an off-the-shelf component: the parser built by Charniak (2000).

Our use of Charniak's parser provides for a loose coupling of a lexicalized model of syntax with a generation system, which is far from ideal. Nevertheless, these experiments provide evidence that lexicalized models of syntax can improve the quality of the outputs of statistics-based generators. Our results motivate work aimed at building generation systems that choose between possible renderings of the same meaning using not only n-gram probabilistic models, but lexicalized models of syntax, such as those proposed by Collins (1997) and Charniak (2000).

2 Statistical Summarization

2.1 Document Compression

To assess the impact that lexicalized models of syntax may have on the task of summary generation, we used a noisy-channel document compression system (Daumé III and Marcu, 2002), which generalizes the sentence compression system developed by Knight and Marcu (2000) to the document level.

In Daumé and Marcu's system, possible document compressions are packed into a shared-forest structure which contains explicit channel-model probability scores. These channel probabilities are based on discourse PCFG probabilities and syntactic PCFG probabilities, as well as compression probabilities. None of these probabilities are lexicalized.

We then use a generic forest ranker which combines these channel probabilities with bigram-based source model probabilities (Langkilde, 2000) to extract the top scoring compression. The only point at which lexicalization enters the model is in the bigram-based source model. This leads to the generation of many poor syntactic structures. For instance, the model does not know the difference between transitive and intransitive verbs, and therefore will often "compress off" the objects of transitive verbs.

After the system was constructed, we evaluated its performance on 16 documents chosen from the Wall Street Journal portion of the Penn Treebank, each containing between 41 and 87 words[2]. We also evaluated it on 5 documents selected from the Mitre corpus (Hirschman et al., 1999), each document containing between 64 and 91 words. We used two corpora to see whether the system's performance varied with text genre. In the evaluations, the system outperformed a baseline system presented by Knight & Marcu (2000), applied iteratively. However, it still made many errors which would not have been made, had it had a good grasp of English grammar. We performed an exhaustive error analysis on the system to see where it could be improved.

2.2 Error Analysis

After analyzing the errors the system made, we classified them into ten classes, each listed in Table 1 with an example of the error and an example of this error fixed.

We then tabulated the frequency with which these error occurred in the summaries produced by the system. They are shown in Table 2.

Three of these error classes can be considered grammaticality errors: DET, MOD, COMP, NOVERB and NUM. These three alone account for 37 errors out of a total of 57 errors. It seemed that the integration of a lexicalized model of syntax into the source model would easily remove all of these problems. Since the document compression system was able to output an n-best list of possible compressions, we were able to rerank this list using Charniak's parser.

2.3 Syntax-Based Reranking

To do this reranking, we modified Charniak's (2000) parser so that instead of outputting the optimal parse tree for a given sentence, it outputs its non-normalized maximum entropy score. We then ran the modified parser on the 1000 best compressions according to the bigram model, normalized the probabilities by length and chose the single best compression.

[1] We are grateful to Eugene Charniak for suggesting these experiments to us.

[2] Because there are an exponential number of summaries that can be generated for any text, the decoder runs out of memory for longer documents; therefore, we selected shorter subtexts from the original documents.

Grammatical Errors

DET - A noun or noun phrase is missing a determiner.

Erroneous "El Paso owns and operates refinery."
Fixed "El Paso owns and operates *a* refinery."

COMP - The complement of a verb or noun is missing, rendering the output ungrammatical.

Erroneous "Banco Exterior was run by politicians who lacked the skills or the will."
Fixed "Banco Exterior was run by politicians who lacked the skills or the will *to do* . . . ."

NUM - Often if the original document contains a percentage (for instance "5%"), the number will be dropped without the percentage sign.

Erroneous "The rate is %.".
Fixed "The rate is 5%.".

NOVERB - Sentence lacking a verb.

Erroneous "Stewart, the builder."
Fixed "*It is named for* Stewart, the builder."

Discourse/Coherence-specific Errors

ANTE - The antecedent of an anaphor has been dropped, resulting in incoherence.

Erroneous "Terms are subject to change, the company said."
Fixed "Terms are subject to change, Banko Exterior said."

CUE+ - Uninterpretable cue-word. For instance, if D_1 and D_2 form a discourse constituent and D_2 is contrasting D_1 and D_2 begins with "But", but D_1 is dropped, so should be "But."

Erroneous "But the proposed transaction calls for an exchange of the debt . . . "
Fixed "The proposed transaction calls for an exchange of the debt . . . "

CUE- - A cue word or phrase is missing, rendering the output incoherent.

Erroneous "The President of the United States urged the armed forces to advance. His commanders did not have the initiative."
Fixed "President of the United States urged the armed forces to advance. *When he did,* his commanders did not have the initiative."

Summarization-specific Errors

MOD - A nominal modifier which should not have been dropped has been and significant meaning is lost.

Erroneous "Tons will fill damp barns across the land."
Fixed "Tons *of vegetables* . . . will fill damp barns across the land."

MISS - The compression misses important information.

EXTRA - The compression contains unimportant information.

Table 1: Types of errors in the texts generated by the model.

This is far from an ideal language model. For instance, there is no guarantee that the structure the parser will derive for the sentence will be the same structure the compression model generated. Furthermore, since the parser assumes only one input sentence, the scores produced by the best parse for two different sentences may not be comparable.

2.4 Evaluation

After reranking, we performed the same error analysis as before. The results for the new error analysis are summarized in Table 3.

We can see from the delta row in Table 3 (negative numbers are good), that we removed most grammaticality errors. We saw modest improvement with respect to the dropping of important modifiers; this is to be expected, though, since the syntax model has no idea of "importance." The same argument explains the minimal change in the missing antecedent problems. We removed all instances of extra information but added seven additional counts of missing information.

Any summarization system must balance length of summary against informational and grammatical quality. In order to have more documents of higher grammaticality, more words are often nec-

	Grammaticality					Discourse			Summarization	
	DET	MOD	COMP	NUM	NOVERB	ANTE	CUE+	CUE-	MISS	EXTRA
wsj_0607	1	1								
wsj_0616	1		1						1	1
wsj_0632									1	
wsj_0654			1			1	1			
wsj_0655			1							
wsj_0667	1	2		1						
wsj_0689			1			1			1	1
wsj_1126		2					1		1	
wsj_1146		1								
wsj_1189	1	2			1					
wsj_1307	1	1	2					1		
wsj_1331									1	
wsj_1346			2						1	
wsj_1376		1	1					1		
wsj_1380	1		1							
wsj_2386			1						1	1
rm5-10	1					1			1	
rm5-22	1		1			2				
rm5-27		1	1							
rm5-6	1				1					
rm5-9										
Count	10	11	13	1	2	5	2	2	8	3

Table 2: Tabulation of the errors in the texts generated by the model.

	Grammaticality					Discourse			Summarization	
	DET	MOD	COMP	NUM	NOVERB	ANTE	CUE+	CUE-	MISS	EXTRA
wsj_0607										
wsj_0616		1							1	
wsj_0632			1						1	
wsj_0654									1	
wsj_0655		1						1	1	
wsj_0667				1						
wsj_0689		2						1	1	
wsj_1126									1	
wsj_1146		1								
wsj_1189										
wsj_1307									1	
wsj_1331		1							1	
wsj_1346		1								
wsj_1376								1	1	
wsj_1380		1						1		
wsj_2386									1	
rm5-10	1								1	
rm5-22			1						1	
rm5-27			1						2	
rm5-6										
rm5-9			1						1	
Count	1	8	4	1	0	4	0	0	15	0
Delta	-9	-3	-9	0	-2	-1	-2	-2	$+7$	-3

Table 3: Tabulation of the errors in the texts generated by the model with syntax-based rescoring.

essary, which reduces the amount of information which can be packed into a summary of comparable length. Here, by removing most of the grammaticality errors, we caused the system to drop some of the important information.

To determine whether the changes in the system are noticeable to a user, we carried out a subjective evaluation. We presented 8 human judges with the outputs generated by the original text compression system, the results after rescoring, and human-generated compressions. These judges were asked to rank outputs on a scale from 1 to 5 (5 being the

	WSJ Texts				Mitre Texts			
	Cmp	Grm	Coh	Qual	Cmp	Grm	Coh	Qual
Old	0.47	3.11	2.98	2.55	0.47	3.57	2.90	2.80
Rescored	0.42	3.41	3.11	2.64	0.30	3.00	3.05	2.23
Hand	0.59	4.38	4.33	3.97	0.46	4.70	4.45	4.10

Table 4: Evaluation Results

best) on metrics of compression rate (Cmp), Grammaticality (Grm), Coherence (Coh) and Compression Quality (Qual). The results of this evaluation are summarized in Table 4.

In the Wall Street Journal data, there was a moderate improvement in grammaticality, coherence and quality, as error analysis suggested. In the Mitre data, grammaticality and quality went down significantly, while coherence remained steady. This can be attributed to two factors. First, there were few errors in the Mitre data to start with and thus less room for improvement; Second, the Mitre data is out of domain for both the document compression system and for the parser, which leads to less reliable statistics.

3 Machine Translation

For our machine translation experiments, we use the statistical MT system of Yamada and Knight (2001). This system produces English translations of foreign language sentences by exploiting three components:

- A Translation model (TM). For any given pair ⟨English parse tree e, foreign language string f⟩, this model returns a probability $P(f|e)$.
- A Language model (LM). For any English tree e, this model returns a probability $P(e)$.
- A Search algorithm. Given a foreign language sentence f, this algorithm searches for the English tree e that maximizes $P(e|f) \approx P(e) \cdot P(f|e)$.

Yamada and Knight (2001; 2002) describe the TM and the search algorithm, respectively. The search algorithm takes a foreign language sentence and produces a vast number of candidate English trees, packed into a forest structure, as in summarization (see Section 2), then searches for the highest-scoring tree. The current algorithm uses a trigram LM, ignoring the internal structure of the candidate trees. Therefore, the system does not necessarily produce syntactically correct translations.

To see the effect of a lexicalized syntax language model, we performed what automatic speech recognition (ASR) researchers call "a cheating experiment". For a given acoustic signal, ASR researchers know both the correct target transcription, A, (which was done by hand) and the current automatic system transcription, B. The probabilistic score for B will be greater than that for A. However, a new knowledge source may provide additional scores that cause A to be reranked higher. This is cheating for two reasons: (1) it does not include a search algorithm that integrates the new knowledge source, and (2) there may be an incorrect string C that scores higher than both A and B under reranking.

This kind of experiment is not regularly done in machine translation because there is no single correct translation A. Using a human translation as a target A does not work, because human translations are often non-literal and current statistical models do not recognize them as good translations. To remedy this, we manually created a set of target translations, which we called "hope" translations. These are good translations that we believe to be within reach of the system: we can reasonably hope that the system would prefer them.

In our experiment, we score both hope sentences, A, and current system translations, B, over a number of examples, using combinations of these knowledge sources:

- T: translation model
- R: word-trigram language model
- C: score from Charniak's parser

Table 5 shows the results. A sentence marked "dec1" is a decoder output (Sentence B) and "hope" is a hope sentence (Sentence A); lower scores are better. The last row shows the average difference of the score (a positive difference means that the system prefers "hope" sentences over current system outputs). Just above the last row is the number of sentences which ranked better.[3]

[3]The score from Charniak's parser (C) is a $-log$(un-normalized prob), thus it may yield negative value. The TM scores are calculated from parse trees, not from

			T	R	C	T+R	T+2R	T+C	T+R+C	T+2C	T+R+2C
1	dec1	great attention his first visit to china	**36.36**	**39.63**	17.44	**75.99**	**115.62**	**53.80**	**93.43**	71.24	110.87
	hope	his first visit is highly important to the chinese side	66.74	54.17	**-5.14**	120.91	175.08	61.60	115.77	**56.46**	**110.63**
2	dec1	he briefed reporters declaring the major contents	**45.47**	**52.62**	16.53	**98.09**	**150.72**	**62.00**	**114.62**	78.53	131.15
	hope	he declared the main contents to the press agency	64.81	56.69	**0.72**	121.50	178.20	65.52	122.22	**66.24**	**122.93**
3	dec1	industrial production and marketing join basic level restored to normal state	**50.84**	**74.34**	25.62	**125.17**	**199.51**	**76.45**	**150.79**	**102.07**	**176.41**
	hope	the basic link between supply and marketing in industry resumed normal status	81.84	83.53	**15.01**	165.37	248.90	96.85	180.38	111.86	195.39
4	dec1	renminbi exchange rate stability this reputation	**16.99**	**43.38**	36.31	**60.37**	**103.75**	53.30	96.68	89.61	132.99
	hope	the renminbi exchange rate continued to maintain stability	27.34	45.43	**18.27**	72.77	118.20	**45.62**	**91.05**	**63.89**	**109.32**
5	dec1	order form is really “ agricultural ” the rapid development	**27.32**	**57.36**	25.96	**84.68**	**142.04**	**53.29**	**110.65**	**79.25**	**136.61**
	hope	third , “ agriculture to order ” has developed rapidly	47.61	62.06	**17.79**	109.67	171.73	65.40	127.46	83.19	145.25
6	dec1	talks have topic hot spot	**23.65**	49.54	28.26	**73.19**	**122.73**	**51.91**	101.45	80.17	129.71
	hope	talks will have many hot topics	39.73	**47.08**	**13.31**	86.81	133.88	53.03	**100.11**	**66.34**	**113.42**
7	dec1	reminding one person alone can not treat	**39.80**	**42.88**	**16.20**	**82.69**	**125.57**	**56.01**	**98.89**	**72.21**	**115.09**
	hope	to remind chen that one-man rule is unacceptable	44.26	73.31	39.15	117.56	190.87	83.41	156.71	122.55	195.86
8	dec1	bad now like this ?	45.15	30.03	7.01	75.18	105.21	52.16	82.19	59.17	89.21
	hope	now it is not good ?	**39.78**	**29.43**	**-10.12**	**69.20**	**98.63**	**29.66**	**59.09**	**19.54**	**48.97**
9	dec1	stratify bimonthly the masses	**17.69**	**32.87**	34.18	**50.56**	**83.43**	51.87	84.74	86.05	118.92
	hope	the masses are stratified	26.52	34.07	**11.30**	60.58	94.65	**37.82**	**71.89**	**49.12**	**83.19**
10	dec1	undercompensation conditions for accept the us side to this aspect	**51.95**	**53.75**	38.74	**105.70**	**159.45**	90.69	**144.44**	129.43	183.18
	hope	the us side did not accept the dprk side ’s proposed compensation terms	52.35	70.89	**27.98**	123.24	194.13	**80.33**	151.22	**108.31**	**179.19**
11	dec1	this is extremely absurd	**21.93**	**18.81**	**2.01**	**40.74**	**59.54**	**23.94**	**42.74**	**25.95**	**44.75**
	hope	this is extremely absurd	24.15	**18.81**	**2.01**	42.96	61.76	26.16	44.96	28.16	46.97
12	dec1	rights and obligations are associated with each other	40.96	**32.37**	**-3.53**	**73.33**	**105.69**	**37.44**	**69.80**	**33.91**	**66.27**
	hope	but rights and obligations are mutually linked	**35.73**	52.06	12.32	87.79	139.84	48.05	100.10	60.37	112.42
13	dec1	mass various channels to reflect the views	**39.88**	**39.22**	16.68	**79.10**	**118.32**	56.55	**95.77**	73.23	112.45
	hope	there are many such channels to reflect the views of the masses	57.78	45.56	**-4.22**	103.33	148.89	**53.56**	99.11	**49.34**	**94.89**
14	dec1	hong kong is our home	**13.93**	**23.39**	**12.34**	**37.32**	**60.71**	**26.27**	**49.66**	**38.61**	**61.99**
	hope	hong kong is our home	16.78	**23.39**	**12.34**	40.17	63.55	29.12	52.50	41.45	64.84
15	dec1	he said china evermore taste	**31.68**	**32.81**	27.96	**64.49**	**97.31**	59.64	**92.45**	87.59	120.41
	hope	he said he often eats chinese food	47.46	54.21	**8.63**	101.67	155.88	**56.09**	110.30	**64.72**	**118.93**
16	dec1	hu jintao said	**5.52**	**14.57**	**39.17**	**20.09**	**34.66**	**44.69**	**59.26**	**83.86**	**98.43**
	hope	hu jintao said	7.90	**14.57**	**39.17**	22.47	37.04	47.07	61.64	86.24	100.81
17	dec1	this is not possible	22.03	**17.21**	**-3.10**	**39.24**	**56.44**	**18.93**	**36.14**	**15.83**	**33.04**
	hope	“ this is not possible	**21.26**	19.97	4.00	41.23	61.20	25.26	45.23	29.27	49.24
18	dec1	this globalization of production	**18.67**	**25.89**	10.68	**44.56**	**70.46**	29.35	55.24	40.03	65.92
	hope	this is the globalization of production	22.08	25.90	**-0.59**	47.98	73.88	**21.49**	**47.39**	**20.90**	**46.80**
19	dec1	many successes possible ?	**24.15**	**32.48**	13.53	**56.63**	**89.11**	37.68	70.16	51.21	83.69
	hope	how much success is possible ?	28.77	35.91	**-4.54**	64.68	100.59	**24.23**	**60.14**	**19.70**	**55.61**
20	dec1	clinton meeting was held this idea	**42.07**	**40.63**	15.04	**82.69**	**123.32**	**57.11**	**97.73**	**72.15**	**112.78**
	hope	holding this meeting was clinton ’s decision	52.66	49.59	**14.71**	102.24	151.83	67.37	116.95	82.08	131.66
21	dec1	reiterative he explained	**10.10**	**17.77**	23.17	**27.87**	**45.63**	33.27	51.03	56.44	74.20
	hope	he has repeatedly explained	18.72	21.66	**7.40**	40.39	62.05	**26.12**	**47.78**	**33.52**	**55.18**
T	dec1		18	19	6	20	20	12	14	9	9
	hope		3	5	18	1	1	9	7	12	12
ave. diff			-1.048	-0.776	0.956	-1.825	-2.601	-0.092	-0.868	0.864	0.088

Table 5: Results of the experiment on the MT system

```
(H37 :ADJUNCT "earlier"
     :LOGICAL-SUBJECT (H5 / "company")
     / "announce"
     :ADJUNCT (H34 :ADJUNCT "its"
                   / "plan"))
```

Figure 1: An underspecified input for the sentence, "Earlier the company announced its plans."

As expected, our current system (T+R) almost never ranks the hope sentence higher than the system sentence. If we replace the trigram model with Charniak's parser (T+C), the hope sentence is ranked higher than the originally preferred sentences in 9 cases. These results are slightly better than reranking with the parser (T+R+C). Best results are obtained by assigning a higher weight to the parser score relative to the translation model (T+2C): here, the hope sentence comes out on top in 12 cases.

4 Natural Language Generation

From the sentences in section 23 of the Penn Treebank, inputs to the HALogen generator system were automatically derived and then regenerated(Langkilde-Geary, 2002). The inputs derived were feature-value dependency structures, where the features represent syntactic relationships between values, and the values were either words in root form or a nested feature-value structure. The inputs were underspecified with respect to properties such as part-of-speech category, tense, voice, and number, as well as constituent order and some closed-class words like auxiliary verbs and determiners. Underspecification tests the ability of a language model to pick the best solution from among a set of choices.

HALogen overgenerates possible expressions for an input, in part because of enumerating possible choices for underspecified details. It then ranks potential outputs using an n-gram model. Both bigram and trigram models were available, but because the trigram model takes two orders of magnitude more time to use (roughly 40 minutes per sentence, versus 1 minute per sentence for the bigram model), the sentences were generated using the bigram model. One hundred sentences were randomly chosen from among the successfully generated outputs. Each was paired with its original Treebank sentence, and then trigram and Charniak-parser scores were calculated for all the sentences. About 3% of the sentence pairs were exact matches. Sentences within pairs had very similar lengths in all cases. The average sentence length was 24 tokens. The original Treebank sentence serves as a gold standard for the generator.

Trigram scores for original Treebank sentences scored better than the output of the generator system 71% of the time. In comparison, Charniak-parse scores preferred the original Treebank sentences 83% of the time. This indicates that the generator system would benefit even more from a statistical model of syntax than from trigrams.

5 Conclusion

In statistical summarization, we have shown that having a lexicalized syntax model reduces the frequency of grammatical errors through both a careful error analysis and a human evaluation. We furthermore show that a lexicalized syntax model might assist in the selection of good translations in a syntax-based machine translation system. Finally, we present results that indicate the importance of lexicalized models of syntax in the HALogen natural language generation system (Langkilde-Geary, 2002).

It seems from these results that the integration of such a language model in these tasks and in statistical natural language generation systems will prove to be fruitful. Most importantly, all three of these systems use the same forest ranking algorithm/component. Thus, if this one component were extended to use a lexicalized model of syntax in place of the current n-gram scoring method, the performance of all three system would undoubtedly improve significantly.

Acknowledgements

This work was partially supported by DARPA-ITO grant N66001-00-1-9814 and a USC Dean Fellowship to Hal Daume III.

sentences. For "dec1" sentences, we use the parse tree returned from the decoder to calculate it. For "hope" sentences, we use a parse tree generated from Collins parser (1997) to calculate it. Thus, there may be different T scores for the same sentence.

References

S. Bangalore and O. Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the International Conference on Computational Linguistics (COLING 2000)*, Hong Kong, China.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97)*, pages 598–603, Providence, Rhode Island, July 27–31.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics NAACL–2000*, pages 132–139, Seattle, Washington, April 29 – May 3.

Eugene Charniak. 2001. Immediate-head parsing for language models. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

Ciprian Chelba and Frederick Jelinek. 1998. Exploiting syntactic structure for language modeling. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 225–231, San Francisco, California. Morgan Kaufmann Publishers.

Michael Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL–97)*, pages 16–23, Madrid, Spain, July 7-12.

Hal Daumé III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the Conference of the Association of Computational Linguistics (ACL 2002)*.

L. Hirschman, M. Light, E. Breck, and J. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization — step one: Sentence compression. In *The 17th National Conference on Artificial Intelligence (AAAI–2000)*, pages 703–710, Austin, TX, July 30th – August 3rd.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL'98)*, Montreal, Canada, August.

Irene Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 2nd International Conference on Natural Language Generation (INLG 2002).*, Arden House, NJ, July.

Irene Langkilde. 2000. Forest-based statistical sentence generation. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, Washington, April 30–May 3.

Brian Roark. 2001. Probabilistic top-down parsing and language modelling. In *Computational Linguistics 27(2)*, pages 249–276.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the Conference of the Association of Computational Linguistics (ACL 2001)*.

K. Yamada and K. Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of the Conference of the Association of Computational Linguistics (ACL 2002)*.