Trigger–Pair Predictors in Parsing and Tagging

Ezra Black, Andrew Finch, Hideki Kashioka
ATR Interpreting Telecommunications
Laboratories
2-2 Hikaridai, Seika-cho
Soraku-gun, Kyoto, Japan 619-02
{black,finch,kashioka}@atr.itl.co.jp

Abstract

In this article, we apply to natural language parsing and tagging the device of trigger-pair predictors, previously employed exclusively within the field of language modelling for speech recognition. Given the task of predicting the correct rule to associate with a parse-tree node, or the correct tag to associate with a word of text, and assuming a particular class of parsing or tagging model, we quantify the information gain realized by taking account of rule or tag trigger-pair predictors, i.e. pairs consisting of a "triggering" rule or tag which has already occurred in the document being processed, together with a specific "triggered" rule or tag whose probability of occurrence within the current sentence we wish to estimate. This information gain is shown to be substantial. Further, by utilizing trigger pairs taken from the same general sort of document as is being processed (e.g. same subject matter or same discourse type)—as opposed to predictors derived from a comprehensive general set of English texts—we can significantly increase this information gain.

1 Introduction

If a person or device wished to predict which words or grammatical constructions were about to occur in some document, intuitively one of the most helpful things to know would seem to be which words and constructions occurred within the last half-dozen or dozen sentences of the document. Other things being equal, a text that has so far been larded with, say, mountaineering terms, is a good bet to continue featuring them. An author with the habit of ending sentences with adverbial clauses of confirmation, e.g. "as we all know", will probably keep up that habit as the discourse progresses.

Within the field of language modelling for speech recognition, maintaining a cache of words that have occurred so far within a document, and using this information to alter probabilities of occurrence of particular choices for the word being predicted, has proved a winning strategy (Kuhn et al., 1990). Models using *trigger pairs* of words, i.e. pairs consisting of a "triggering" word which has already occurred in the document being processed, plus a specific "triggered" word whose probability of occurrence as the next word of the document needs to be estimated, have yielded perplexity[1] reductions of 29–38% over the baseline trigram model, for a 5-million-word Wall Street Journal training corpus (Rosenfeld, 1996).

This paper introduces the idea of using trigger-pair techniques to assist in the prediction of rule and tag occurrences, within the context of natural-language parsing and tagging. Given the task of predicting the correct rule to associate with a parse-tree node, or the correct tag to associate with a word of text, and assuming a particular class of parsing or tagging model, we quantify the information gain realized by taking account of rule or tag trigger-pair predictors, i.e. pairs consisting of a "triggering" rule or tag which has already occurred in the document being processed, plus a specific "triggered" rule or tag whose probability of occurrence within the current sentence we wish to estimate.

In what follows, Section 2 provides a basic overview of trigger-pair models. Section 3 describes the experiments we have performed, which to a large extent parallel successful modelling experiments within the field of language modelling for speech recognition. In the first experiment, we investigate the use of trigger pairs to predict both rules and tags over our full corpus of around a million words. The subsequent experiments investigate the

[1] See Section 2.

additional information gains accruing from trigger-pair modelling when we know what sort of document is being parsed or tagged. We present our experimental results in Section 4, and discuss them in Section 5. In Section 6, we present some example trigger pairs; and we conclude, with a glance at projected future research, in Section 7.

2 Background

Trigger-pair modelling research has been pursued within the field of language modelling for speech recognition over the last decade (Beeferman et al., 1997; Della Pietra et al., 1992; Kupiec, 1989; Lau, 1994; Lau et al., 1993; Rosenfeld, 1996).

Fundamentally, the idea is a simple one: if you have recently seen a word in a document, then it is more likely to occur again, or, more generally, the prior occurrence of a word in a document affects the probability of occurrence of itself and other words.

More formally, from an information-theoretic viewpoint, we can interpret the process as the relationship between two dependent random variables. Let the outcome (from the alphabet of outcomes $\mathcal{A}_Y$) of a random variable Y be observed and used to predict a random variable X (with alphabet $\mathcal{A}_X$). The probability distribution of X, in our case, is dependent on the outcome of Y.

The average amount of information necessary to specify an outcome of X (measured in bits) is called its *entropy* $H(X)$ and can also be viewed as a measure of the average ambiguity of its outcome:[2]

$$H(X) = \sum_{x \in \mathcal{A}_X} -P(x) \log_2 P(x) \qquad (1)$$

The *mutual information* between X and Y is a measure of entropy (ambiguity) reduction of X from the observation of the outcome of Y. This is the entropy of X minus its *a posteriori* entropy, having observed the outcome of Y.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \end{aligned} \qquad (2)$$

The dependency information between a word and its history may be captured by the *trigger pair*.[3] A trigger pair is an ordered pair of words t and w. Knowledge that the trigger word t has occurred within some *window* of words in the history, changes the probability estimate that word w will occur subsequently.

Selection of these triggers can be performed by calculating the average mutual information between word pairs over a training corpus. In this case, the alphabet $\mathcal{A}_X = \{w, \overline{w}\}$, the presence or absence of word w; similarly, $\mathcal{A}_Y = \{t, \overline{t}\}$, the presence or absence of the triggering word in the history.

This is a measure of the effect that the knowledge of the occurrence of the triggering word t has on the occurence of word w, in terms of the entropy (and therefore perplexity) reduction it will provide. Clearly, in the absence of other context (i.e. in the case of the *a priori* distribition of X), this information will be additional. However, once related contextual information is included (for example by building a trigram model, or, using other triggers for the same word), this is no longer strictly true.

Once the trigger pairs are chosen, they may be used to form constraint functions to be used in a maximum-entropy model, alongside other constraints. Models of this form are extremely versatile, allowing the combination of short- and long-range information. To construct such a model, one transforms the trigger pairs into *constraint functions* $f(t,w)$:

$$f(t,w) = \begin{cases} 1 & \text{if } t \in \text{history and} \\ & \text{next word} = w \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

The expected values of these functions are then used to constrain the model, usually in combination of with other constraints such as similar functions embodying uni-, bi- and trigram probability estimates.

(Beeferman et al., 1997) models more accurately the effect of distance between triggering and triggered word, showing that for non-self-triggers,[4] the triggering effect decays exponentially with distance. For self-triggers,[5] the effect is the same except that the triggering effect is lessened within a short range of the word. Using a model of these distance effects, they are able to improve the performance of a trigger model.

We are unaware of any work on the use of trigger pairs in parsing or tagging. In fact, we have not found any previous research in which extrasentential data of any sort are applied to the problem of parsing or tagging.

3 The Experiments

3.1 Experimental Design

In order to investigate the utility of using long-range trigger information in tagging and parsing

[2] A more intuitive view of entropy is provided through *perplexity* (Jelinek et al., 1977) which is a measure of the number of choices, on average, there are for a random variable. It is defined to be: $2^{H(X)}$.

[3] For a thorough description of trigger-based modelling, see (Rosenfeld, 1996).

[4] i.e. words which trigger words other than themselves

[5] i.e. words which trigger themselves

tasks, we adopt the simple mutual-information approach used in (Rosenfeld, 1996). We carry over into the domain of tags and rules an experiment from Rosenfeld's paper the details of which we outline below.

The idea is to measure the information contributed (in bits, or, equivalently in terms of perplexity reduction) by using the triggers. Using this technique requires special care to ensure that information "added" by the triggers is indeed additional information.

For this reason, in all our experiments we use the unigram model as our base model and we allow only one trigger for each tag (or rule) token.[6] We derive these unigram probabilities from the training corpus and then calculate the total mutual information gained by using the trigger pairs, again with respect to the training corpus.

When using trigger pairs, one usually restricts the trigger to occur within a certain window defined by its distance to the triggered token. In our experiments, the window starts at the sentence prior to that containing the token and extends back W (the window size) sentences. The choice to use sentences as the unit of distance is motivated by our intention to incorporate triggers of this form into a probabilistic treebank-based parser and tagger, such as (Black et al., 1998; Black et al., 1997; Brill, 1994; Collins, 1996; Jelinek et al., 1994; Magerman, 1995; Ratnaparkhi, 1997). All such parsers and taggers of which we are aware use only intrasentential information in predicting parses or tags, and we wish to remove this information, as far as possible, from our results [7]. The window was not allowed to cross a document boundary. The perplexity of the task before taking the trigger-pair information into account for tags was 224.0 and for rules was 57.0.

The characteristics of the training corpus we employ are given in Table 1. The corpus, a subset[8] of the ATR/Lancaster General-English Treebank (Black et al., 1996), consists of a sequence of sentences which have been tagged and parsed by human experts in terms of the ATR English Grammar, a broad-coverage grammar of English with a high level of analytic detail (Black et al., 1996; Black et al., 1997). For instance, the tagset is both semantic and syntactic, and includes around 2000 different tags, which classify nouns, verbs, adjectives and adverbs via over 100 semantic categories. As examples of the level of syntactic detail, exhaustive syntactic and semantic analysis is performed on all nominal compounds; and the full range of attachment sites is available within the Grammar for sentential and phrasal modifiers, and are used precisely in the Treebank. The Treebank actually consists of a set of documents, from a variety of sources. Crucially for our experiments (see below), the idea[9] informing the selection of (the roughly 2000) documents for inclusion in the Treebank was to pack into it the maximum degree of document variation along many different scales---document length, subject area, style, point of view, etc.---but without establishing a single, predetermined classification of the included documents.

1868 documents
80299 sentences
904431 words (tag instances)
1622664 constituents (rule instances)
1873 tags utilized
907 rules utilized
11.3 words per sentence, on average

Table 1: Characteristics of Training Set (Subset of ATR/Lancaster General-English Treebank)

In the first experiment, we examine the effectiveness of using trigger pairs over the entire training corpus. At the same time we investigate the effect of varying the window size. In additional experiments, we observe the effect of partitioning our training dataset into a few relatively homogeneous subsets, on the hypothesis that this will decrease perplexity. It seems reasonable that in different text varieties, different sets of trigger pairs will be useful, and that tokens which do not have effective triggers within one text variety may have them in another.[10]

To investigate the utility of partitioning the dataset, we construct a separate set of trigger pairs for each class. These triggers are only active for their respective class and are independent of each other. Their total mutual information is compared to that derived in exactly the same way from a random partition of our corpus into the same number of classes, each comprised of the same number of documents.

Our training data partitions naturally into four subsets, shown in Table 2 as Partitioning 1 ("Source"). Partitioning 2, "List Structure", puts all documents which contain at least some HTML-like "List" markup (e.g. LI (=List Item))[11] in one

[6] By rule assignment, we mean the task of assigning a rule-name to a node in a parse tree, given that the constituent boundaries have already been defined.

[7] This is not completely possible, since correlations, even if slight, will exist between intra- and extrasentential information

[8] specifically, a roughly-900,000-word subset of the full ATR/Lancaster General-English Treebank (about 1.05 million words), from which all 150,000 words were excluded that were treebanked by the two least accurate ATR/Lancaster treebankers (expected hand-parsing error rate 32%, versus less than 10% overall for the three remaining treebankers)

[9] see (Black et al., 1996)

[10] Related work in topic-specific trigram modelling (Lau, 1994) has led to a reduction in perplexity.

[11] All documents in our training set are marked up in HTML-like annotation.

subset, and all other documents in the other subset. By merging Partitionings 1 and 2 we obtain Partitioning 3, "Source Plus List Structure". Partitioning 4 is "Source Plus Document Type", and contains 9 subsets, e.g. "Letters; diaries" (subset 8) and "Novels; stories; fables" (subset 7). With 13 subsets, Partitioning 5, "Source Plus Domain", includes e.g. "Social Sciences" (subset 9) and Recreation (subset 1). Partitionings 4 and 5 were effected by actual inspection of each document, or at least of its title and/or summary, by one of the authors. The reason we included Source within most partitionings was to determine the extent to which information gains were additive.[12]

4 Experimental Results

4.1 Window Size

Figure 1 shows the effect of varying the window size from 1 to 500 for both rule and tag tokens. The optimal window size for tags was approximately 12 sentences (about 135 words) and for rules it was approximately 6 sentences (about 68 words). These values were used for all subsequent experiments. It is interesting to note that the curves are of similar shape for both rules and tags and that the optimal value is not the largest window size. Related effects for words are reported in (Lau, 1994; Beeferman et al., 1997). In the latter paper, an exponential model of distance is used to penalize large distances between triggering word and triggered word. The variable window used here can be seen as a simple alternative to this.

One explanation for this effect in our data is, in the case of tags, that topic changes occur in documents. In the case of rules, the effect would seem to indicate a short span of relatively intense stylistic carryover in text. For instance, it may be much more important, in predicting rules typical of list structure, to know that similar rules occurred a few sentences ago, than to know that they occurred dozens of sentences back in the document.

4.2 Class-Specific Triggers

Table 3 shows the improvement in perplexity over the base (unigram) tag and rule models for both the randomly-split and the hand-partitioned training sets. In every case, the meaningful split yielded significantly more information than the random split. (Of course, the results for randomly-split training sets are roughly the same as for the unpartitioned training set (Figure 1)).

[12]For instance, compare the results for Partitionings 1, 2, and 3 in this regard.

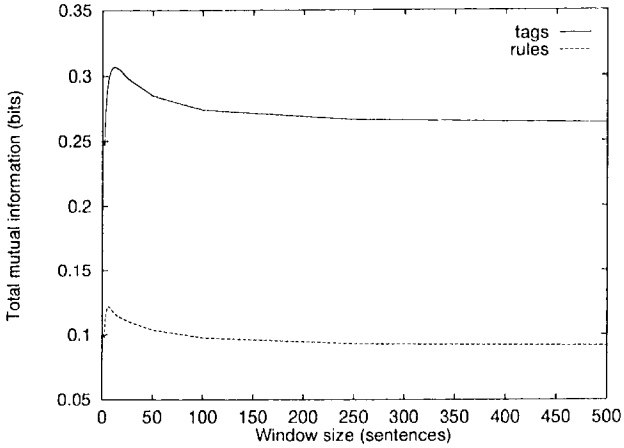

Figure 1: Mutual information gain varying window size

5 Discussion

The main result of this paper is to show that analogous to the case of words in language modelling, a significant amount of extrasentential information can be extracted from the long-range history of a document, using trigger pairs for tags and rules. Although some redundancy of information is inevitable, we have taken care to exclude as much information as possible that is already available to (intrasentential-data-based, i.e. all known) parsers and taggers.

Quantitatively, the studies of (Rosenfeld, 1996) yielded a total mutual information gain of 0.38 bits, using Wall Street Journal data, with one trigger per word. In a parallel experiment, using the same technique, but on the ATR/Lancaster corpus, the total mutual information of the triggers for *tags* was 0.41 bits. This figure increases to 0.52 bits when tags further away than 135 tags (the approximate equivalent in words to the optimal window size in sentences) are excluded from the history. For the remainder of our experiments, we do not use as part of the history the tags/rules from the sentence containing the token to be predicted. This is motivated by our wish to exclude the intrasentential information which is already available to parsers and taggers.

In the case of tags, using the optimal window size, the gain was 0.31 bits, and for rules the information gain was 0.12 bits. Although these figures are not as large as for the case where intrasentential information is incorporated, they are sufficiently close to encourage us to exploit this information in our models.

For the case of words, the evidence shows that triggers derived in the same manner as the triggers in our experiments, can provide a substantial amount of new information when used in combination with sophisticated language models. For example, (Rosenfeld, 1996) used a maximum-entropy

Part. 1: *Source*		Part. 4: *Source + Doc Type*		Part. 5: *Source + Domain*	
Class Name	Sents	Class Name	Sents	Class Name	Sents
1: Assoc. Press, WSJ	8851	1: Legislative (incl. *Srce*.2)	5626	1: Recreation	3545
2: Canadian Hansards	5002	2: Transcripts (incl. *Srce*.4)	44287	2: Business	2055
3: General English	23105	3: News (incl. most *Srce*.1)	8614	3: Science, Techn.	4018
4: Travel-domain dialgs	43341	4: Polemical essays	5160	4: Humanities	2224
Part. 2: *List Structure*		5: Reports; FAQs; listings	11440	5: Daily Living	896
Class Name	Sents	6: Idiom examples	666	6: Health, Education	1649
1: Contains lists	14147	7: Novels; stories; fables	741	7: Government, Polit.	1768
2: Contains no lists	66152	8: Letters; diaries	1997	8: Travel	2667
Part. 3: *Source + List Structure*		9: Legal cases; constitutions	1768	9: Social Sciences	3617
Class Name	Sents			10: Idiom examp. sents	666
1: Assoc. Press, WSJ	8851			11: Canadian Hansards	5002
2: Canadian Hansards	5002			12: Assoc. Press, WSJ	8851
3: Contains lists (Gen.)	11998			13: Travel dialgs	43341
4: Contains no lists (Gen.)	11117				
5: Travel-domain dialogues	43341				

Table 2: Training Set Partitions

Partitioning	Perplexity reduction for tags		Perplexity reduction for rules	
	Meaningful partition	Random	Meaningful partition	Random
1: *Source*	28.40%	16.66%	15.44%	6.30%
2: *List Structure*	20.39%	18.71%	10.55%	7.46%
3: *Source Plus List Structure*	28.74%	17.12%	15.61%	6.50%
4: *Source Plus Document Type*	30.11%	18.15%	16.20%	6.82%
5: *Source Plus Domain*	31.55%	19.39%	16.60%	7.34%

Table 3: Perplexity reduction using class-specific triggers to predict tags and rules

#	Triggering Tag	Triggered Tag	I.e. Words Like These:	Trigger Words Like These:
1	NP1LOCNM	NP1STATENM	Hill, County, Bay, Lake	Utah, Maine, Alaska
2	JJSYSTEM	NP1ORG	national, federal, political	Party, Council, Department
3	IIDESPITE	CFYET	despite	yet (conjunction)
4	PN1PERSON	LEBUT22	everyone, one, anybody	(not) only, (not) just
5	...	MPRICE	..., ..., ...	$452,983,000, $10,000, $19.95
6	IIAT(SF)	MPHONE22	at (sent.-final, +/-":")	913-3434 (follows area code)
7	IIFROM(SF)	MZIP	from (sent.-final, +/-":")	22314-1698 (postal zipcode)
8	NNUNUM	NN1MONEY	25%, 12", 9.4m3	profit, price, cost

Table 4: Selected Tag Trigger-Pairs, ATR/Lancaster General-English Treebank

#	A Construction Like This:	Triggers A Construction Like This:
1a	Interrupter Phrase -> * Or -	Sentence -> Interrupter P+Phrasal (Non-S)
1b	*Example:* *, -	*Example:* * DIG. AM/FM TUNER
2a	VP -> Verb+Interrupter Phrase+Obj/Compl	Interrupter Phrase -> ,+Interrupter+,
2b	*Example:* starring--surprise, surprise--men	*Example:* , according to participants ,
3a	Noun Phrase -> Simple Noun Phrase+Num	Num -> Num +PrepP with Numerical Obj
3b	*Example:* Lows around 50	*Example:* (Snow level) 6000 to 7000
4a	Verb Phrase -> Adverb Phrase+Verb Phrase	Auxiliary VP -> Model/Auxilliary Verb+Not
4b	*Example:* just need to understand it	*Example:* do not
5a	Question -> Be+NP+Object/Complement	Quoted Phrasal -> "+Phrasal Constit+"
5b	*Example:* Is it possible?	*Example:* "Mutual funds are back."

Table 5: Selected Rule Trigger-Pairs, ATR/Lancaster General-English Treebank

#	Triggering Tag	Triggered Tag	I.e. Words Like These:	Trigger Words Like These:
1	VVNSEND	NP1STATENM	shipped, distributed	Utah, Maine, Alaska
2	NP1LOCNM	NP1STATENM	Hill, County, Bay, Lake	Utah, Maine, Alaska
For training-set document class Recreation (1) vs. for unpartitioned training set (2)				
3	VVOALTER	NN2SUBSTANCE	inhibit, affect, modify	tumors, drugs, agents
4	JJPHYS-ATT	NN2SUBSTANCE	fragile, brown, choppy	pines, apples, chemicals
For training-set document class Health And Education (3) vs. for unpartitioned training set (4)				
5	NN1TIME	NN2MONEY	period, future, decade	expenses, fees, taxes
6	NP1POSTFRMNM	NN2MONEY	Inc., Associates, Co.	loans, damages, charges
For training-set document class Business (5) vs. for unpartitioned training set (6)				
7	DD1	DDQ	this, that, another, each	which
8	DDQ	DDQ	which	which
For training-set document class Travel Dialogues (7) vs. for unpartitioned training set (8)				

Table 6: Selected Tag Trigger–Pairs, ATR/Lancaster General–English Treebank: Contrasting Trigger–Pairs Arising From Partitioned vs. Unpartitioned Training Sets

model trained on 5 million words, with only trigger, uni–, bi– and trigram constraints, to measure the test–set perpexity reduction with respect to a "compact" backoff trigram model, a well-respected model in the language–modelling field. When the top six triggers for each word were used, test–set perplexity was reduced by 25%. Furthermore, when a more sophisticated version of this model[13] was applied in conjunction with the SPHINX II speech recognition system (Huang et al., 1993), a 10-14% reduction in word error rate resulted (Rosenfeld, 1996). We see no reason why this effect should not carry over to tag and rule tokens, and are optimistic that long–range trigger information can be used in both parsing and tagging to improve performance.

For words (Rosenfeld, 1996), *self–triggers*—words which triggered themselves—were the most frequent kind of triggers (68% of all word triggers were self–triggers). This is also the case for tags and rules. For tags, 76.8% were self-triggers, and for rules, 96.5% were self–triggers. As in the case of words, the set of self–triggers provides the most useful predictive information.

6 Some Examples

We will now explicate a few of the example trigger pairs in Tables 4–6. Table 4 Item 5, for instance, captures the common practice of using a sequence of points, e.g. ....., to separate each item of a (price) list and the price of that item. Items 6 and 7 are similar cases (e.g. "contact/call (someone) at:" + phone number; "available from:" + source, typically including address, hence zipcode). These correlations typically occur within listings, and, crucially for their usefulness as triggers, typically occur many at a time.

When triggers are drawn from a relatively homogeneous set of documents, correlations emerge which seem to reflect the character of the text type involved. So in Table 6 Item 5, the proverbial equation of time and money emerges as more central to Business and Commerce texts than the different but equally sensible linkup, within our overall training set, between business corporations and money.

Turning to rule triggers, Table 5 Item 1 is more or less a syntactic analog of the tag examples Table 4 Items 5–7, just discussed. What seems to be captured is that a particular style of listing things, e.g. * + listed item, characterizes a document as a whole (if it contains lists); further, listed items are not always of the same phrasal type, but are prone to vary syntactically. The same document that contains the list item "* DIG. AM/FM TUNER", for instance, which is based on a Noun Phrase, soon afterwards includes "* WEATHER PROOF" and "* ULTRA COMPACT", which are based on Adjective Phrases.

Finally, as in the case of the tag trigger examples of Table 6, text–type–particular correlations emerge when rule triggers are drawn from a relatively homogeneous set of documents. A trigger pair of constructions specific to Class 1 of the Source partitioning, which contains only Associated Press newswire and Wall Street Journal articles, is the following: A sentence containing both a quoted remark and an attribution of that remark to a particular source, triggers a sentence containing simply a quoted remark, without attribution. (E.g. *"The King was in trouble," Wall wrote.* triggers *"This increased the King's bitterness."*.) This correlation is essentially absent in other text types.

[13] trained on 38 million words, and also employing distance–2 N-gram constraints, a unigram cache and a conditional bigram cache (this model reduced perplexity over the baseline trigram model by 32%)

7 Conclusion

In this paper, we have shown that, as in the case of words, there is a substantial amount of information outside the sentence which could be used to supplement tagging and parsing models. We have also shown that knowledge of the type of document being processed greatly increases the usefulness of triggers. If this information is known, or can be predicted accurately from the history of a given document being processed, then model interpolation techniques (Jelinek et al., 1980) could be employed, we anticipate, to exploit this to useful effect.

Future research will concentrate on incorporating trigger-pair information, and extrasentential information more generally, into more sophisticated models of parsing and tagging. An obvious first extention to this work, for the case of tags, will be, following (Rosenfeld, 1996), to incorporate the triggers into a maximum entropy model using trigger pairs in addition to unigram, bigram and trigram constraints. Later we intend to incorporate trigger information into a probabilistic English parser/tagger which is able to ask complex, detailed questions about the contents of a sentence. From the results presented here we are optimistic that the additional, extrasentential information provided by trigger pairs will benefit such parsing and tagging systems.

References

D. Beeferman, A. Berger, and J. Lafferty. 1997. A Model of Lexical Attraction and Repulsion. In *Proceedings of the ACL-EACL'97 Joint Conference*, Madrid.

E. Black, S. Eubank, H. Kashioka, J. Saia. 1998. Reinventing Part-of-Speech Tagging. *Journal of Natural Language Processing (Japan)*, 5:1.

E. Black, S. Eubank, H. Kashioka. 1997. Probabilistic Parsing of Unrestricted English Text, With A Highly-Detailed Grammar. In *Proceedings, Fifth Workshop on Very Large Corpora*, Beijing/Hong Kong.

E. Black, S. Eubank, H. Kashioka, R. Garside, G. Leech, and D. Magerman. 1996. Beyond skeleton parsing: producing a comprehensive large-scale general-English treebank with full grammatical analysis. In *Proceedings of the 16th Annual Conference on Computational Linguistics*, pages 107–112, Copenhagen.

E. Brill. 1994. Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722-727, Seattle, Washington. American Association for Artificial Intelligence.

M. Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Languistics*, Santa Cruz.

S. Della Pietra, V. Della Pietra, R. Mercer, S. Roukos. 1992. Adaptive language modeling using minimum discriminant information. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, I:633-636.

X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K. F. Lee, and R. Rosenfeld. 1993. The SPHINX-II speech recognition system: an overview. *Computer Speech and Language*, 2:137–148.

F. Jelinek, R. L. Mercer, L. R. Bahl, J. K. Baker. 1977. Perplexity—a measure of difficulty of speech recognition tasks. In *Proceedings of the 94th Meeting of the Acoustic Society of America*, Miami Beach, FL.

F. Jelinek and R. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Pattern Recognition In Practice*, E. S. Gelsema and N. L. Kanal, eds., pages 381–402, Amsterdam: North Holland.

F. Jelinek, J. Lafferty, D. Magerman, R. Mercer, A. Ratnaparkhi, S. Roukos. 1994. Decision tree parsing using a hidden derivation model. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265, Plainsboro, New Jersey. Advanced Research Projects Agency.

R. Kuhn, R. De Mori. 1990. A Cache Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

J. Kupiec. 1989. Probabilistic models of short and long distance word dependencies in running text. In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 290–295.

R. Lau, R. Rosenfeld, S. Roukos. 1993. Trigger-based language models: a maximum entropy approach. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, II:45–48.

R. Lau. 1994. Adaptive Statistical Language Modelling. *Master's Thesis*, Massachusetts Institute of Technology, MA.

D. Magerman. 1995. Statistical decision-tree models for parsing. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, Massachusetts. Association for Computational Linguistics.

A. Ratnaparkhi. 1997. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. In *Proceedings, Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech and Language*, 10:187–228.