

THE ROYAL
SOCIETY

PHILOSOPHICAL
TRANSACTIONS ŏ

**MATHEMATICAL,
PHYSICAL**
& ENGINEERING
SCENGINEERING

THE ROYAL

PHILOSOPHICAL
TRANSACTIONS ŏ

statistical language models Incorporating linguistic structure into

THE ROYAL
SOCIETY

Ronald Rosenfeld

 $OF -$

doi: 10.1098/rsta.2000.0588 Phil. Trans. R. Soc. Lond. A 2000 **358**, 1311-1324

Email alerting service

the article or click **[here](http://rsta.royalsocietypublishing.org/cgi/alerts/ctalert?alertType=citedby&addAlert=cited_by&saveAlert=no&cited_by_criteria_resid=roypta;358/1769/1311&return_type=article&return_url=http://rsta.royalsocietypublishing.org/content/358/1769/1311.full.pdf)** article - sign up in the box at the top right-hand corner of Receive free email alerts when new articles cite this

MATHEMATIC PHYSICAL & ENGINEERI SCIENCES

<http://rsta.royalsocietypublishing.org/subscriptions> To subscribe to Phil. Trans. R. Soc. Lond. A go to:

PHILOSOPHICAL
TRANSACTIONS

ATHEMATICAL

NYA

H

PHILOSOPHICAL
TRANSACTIONS

Incorporating linguistic structure into
statistical language models porating linguistic structure in
statistical language models statistical language models
BY RONALD ROSENFELD

BY RONALD ROSENFELD
School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA 15919, USA *Pittsburgh, PA 15213, USA*

Futsburgh, FA 19215, OSA
Statistical language models estimate the distribution of natural language for the
purpose of improving various language technology applications. Ironically, the most Statistical language models estimate the distribution of natural language for the purpose of improving various language technology applications. Ironically, the most successful models of this type take little advantage of Statistical language models estimate the distribution of natural language for the purpose of improving various language technology applications. Ironically, the most successful models of this type take little advantage of purpose of improving various language technology applications. Ironically, the most successful models of this type take little advantage of the nature of language. I review the extent to which various aspects of natural la successful models of this type take little advantage of the nature of language. I
review the extent to which various aspects of natural language are captured in current
models. I then describe a general framework, recently review the extent to which various aspects of natural language are captured in current
models. I then describe a general framework, recently developed at our laboratory, for
incorporating arbitrary linguistic structure int models. I then describe a general framework, recently developed at our laboratory, for
incorporating arbitrary linguistic structure into a statistical framework, and present
a methodology for eliciting linguistic features incorporating arbitrary linguistic structure into a statistical framework, and present
a methodology for eliciting linguistic features currently missing from the model.
Finally, I ponder our failure heretofore to integrate a methodology for eliciting linguistic features
Finally, I ponder our failure heretofore to integrate
framework, and suggest possible reasons for it. framework, and suggest possible reasons for it.
Keywords: statistical language modelling;

human language technologies; feature induction

1. Introduction

1. Introduction
Statistical language models (SLMs) estimate the probability of sentences in natural
language using large amounts of training data. SLMs are used in a variety of language Statistical language models (SLMs) estimate the probability of sentences in natural
language using large amounts of training data. SLMs are used in a variety of language
technology applications, such as speech recognition, Statistical language models (SLMs) estimate the probability of sentences in natural
language using large amounts of training data. SLMs are used in a variety of language
technology applications, such as speech recognition, language using large amounts of training data. SLMs are used in a variety of language technology applications, such as speech recognition, document classification, optical character recognitions, machine translation, and m technology applications, such as speech recognition, document classification, optical character recognitions, machine translation, and more. In speech recognition, for example, an incoming acoustic signal a is given. The character recognitions, machine transla
example, an incoming acoustic signal a
that maximizes the posterior $P(s | a)$: s = arg max_s $P(s \mid a)$ = arg max_s $P(a \mid s) \cdot P(s)$; (1.1)

$$
s^* = \arg \max_s P(s \mid a) = \arg \max_s P(a \mid s) \cdot P(s), \tag{1.1}
$$

where the language model $P(s)$ plays the role of the prior.

here the language model $P(s)$ plays the role of the prior.
A given language model M is often evaluated by its *perplexity*,

$$
M \text{ is often evaluated by its } perpetuity,
$$
\n
$$
\text{perplexity}(M) = 2^{H(P;P_M)},\tag{1.2}
$$

perplexity(M) = $2^{H(P;P_M)}$, (1.2)
where $H(P; P_M)$ is the cross entropy between the distribution P_M described by the
model and P_D the true distribution of the data where $H(P; P_M)$ is the cross entropy between the
model and P_D , the true distribution of the data.
Ironically the most successful SLM techniques Ironically, the most successful SLM techniques use very little knowledge of what
Ironically, the most successful SLM techniques use very little knowledge of what
Ironically, the most successful SLM techniques use very lit

model and P_D , the true distribution of the data.
Ironically, the most successful SLM techniques use very little knowledge of what
language really is. Attempts to incorporate linguistic theories or even linguistic intu-
 Ironically, the most successful SLM techniques use very little knowledge of what
language really is. Attempts to incorporate linguistic theories or even linguistic intu-
ition into SLMs have met with very limited success. language really is. Attempts to incorporate linguistic theories or even linguistic intuition into SLMs have met with very limited success. In what follows, $\S 2$ lists various aspects of natural language, and reviews the ition into SLMs have met with very limited success. In what follows, $\S 2$ lists various aspects of natural language, and reviews the extent to which they are captured in current models. Section 3 describes a general fram aspects of natural language, and reviews the extent to which they are captured in current models. Section 3 describes a general framework, recently developed at our in $\S 4$ I ponder the SLM community's failure to integrate linguistic theories into a statistical framework, and suggest possible reasons for it. laboratory, for integrating linguistic features into a statistical framework. Finally,

Phil. Trans. R. Soc. Lond. A (2000) 358, 1311-1324

1311

Table 1. *Natural language sentences*

(Example average length sentences from the BN corpus.)

WANDILE ZOTHE DO YOU PERSONALLY KNOW PEOPLE WHO WERE WANDILE ZOTHE DO YOU PERSONALLY KNOW PEOPLE WHO WERE
ARRESTED AND TORTURED DURING THE APARTHEID ERA </s>

MANDILE ZOTILE DO TOUT ERSOMALLY NINOW THOT LE WHO WERE ARRESTED AND TORTURED DURING THE APARTHEID ERA $\langle s \rangle$
SO HE PROBABLY WILL HAVE TO HAVE THEM TAXED BECAUSE
THEY'RE NOT A TRADITIONAL PENSION FUND $\langle s \rangle$ SO HE PROBABLY WILL HAVE TO HAVE THEM TAXED BECAUSE THEY'RE NOT A TRADITIONAL PENSION FUND \langle /s \rangle

BUT THE TOBACCO COMPANIES AND NASCAR OFFICIALS SAY THEIR FANS ARE WILDLY LOYAL TO RACE ADVERTISERS \langle /s>

THERE ARE A LOT OF QUALITY SWEATERS IN THE MARKET RIGHT NOW CASHMERE AND CASHMERE BLENDS </s>

POLICE SAY THE MAN RAN FROM THE FRONT OF THE HOUSE AND CAME AROUND THIS CORNER ϵ/s POLICE SAY THE MAN RAN FROM THE I
CAME AROUND THIS CORNER </s>

2. Linguistic structure in statistical language models (*a*) *Baseline: the* ⁿ*-gram*

(a) *Baseline: the n-gram*
Almost all language models estimate the probability of a sentence s by using the
chain rule to decompose it into a product of conditional probabilities. *chain rule* to decompose it into a product of conditional probabilities,

$$
\Pr(s) \stackrel{\text{def}}{=} \Pr(w_1, \dots, w_n) = \prod_{i=1}^n \Pr(w_i \mid w_1, \dots, w_{i-1}) \stackrel{\text{def}}{=} \prod_{i=1}^n \Pr(w_i \mid h_i), \quad (2.1)
$$

 $i=1$ $i=1$
where $h_i \stackrel{\text{def}}{=} \{w_1, \ldots, w_{i-1}\}$ is the *history* when predicting word w_i .
The most commonly used language model, the *n*-gram, makes the f

here $h_i \stackrel{\text{def}}{=} \{w_1, \ldots, w_{i-1}\}$ is the *history* when predicting word w_i .
The most commonly used language model, the *n*-gram, makes the further simpli-
ing assumption: where $h_i \stackrel{\text{def}}{=} \{w_1, \ldots$
The most commo
fying assumption: $| w_{i-n+1}, \ldots, w_{i-1} \rangle.$ (2.2)

$$
P(w_i | h_i) \approx P(w_i | w_{i-n+1}, \dots, w_{i-1}). \tag{2.2}
$$

 $P(w_i | h_i) \approx P(w_i | w_{i-n+1}, \dots, w_{i-1}).$ (2.2)
The *n*-gram captures correlations among nearby words reasonably well. Not surpris-
ingly it captures little else. This can be best appreciated by observing 'sentences' The *n*-gram captures correlations among nearby words reasonably well. Not surprisingly, it captures little else. This can be best appreciated by observing 'sentences' senerated from this model. Table 1 lists example sent The *n*-gram captures correlations among nearby words reasonably well. Not surprisingly, it captures little else. This can be best appreciated by observing 'sentences' generated from this model. Table 1 lists example sent ingly, it captures little else. This can be best appreciated by observing 'sentences'
generated from this model. Table 1 lists example sentences from the Broadcast News
(BN) corpus: a corpus of some 13 million sentences tr generated from this model. Table 1 lists example sentences from the Broadcast News
(BN) corpus: a corpus of some 13 million sentences transcribed from TV and radio
news-related programmes between 1992 and 1996 (Graff 1997) \blacktriangleleft (BN) corpus: a corpus of some 13 million sentences transcribed from TV and radio news-related programmes between 1992 and 1996 (Graff 1997). This complete corpus \succ was used to train a state-of-the-art trigram la news-related programmes between 1992 and 1996 (Graff 1997). This complete corpus
was used to train a state-of-the-art trigram language model, which was, in turn, used
in generative mode to produce 'pseudo-sentences', examp was used
in genera
table 2.
It is no able 2.
It is not difficult for people to tell these two language sources apart. In an infor-

table 2.
It is not difficult for people to tell these two language sources apart. In an informal blind study we conducted on Carnegie Mellon's Sphinx speech research group,
classification accuracies of 95% were achieve It is not difficult for people to tell these two language sources apart. In an informal blind study we conducted on Carnegie Mellon's Sphinx speech research group, classification accuracies of 95% were achieved. It is also mal blind study we conducted on Carnegie Mellon's Sphinx speech research group,
classification accuracies of 95% were achieved. It is also easy to appreciate how such
judgements are made, since just about every aspect of n \sim classification accuracies of 95% were achieved. It is also easy to appreciate how such judgements are made, since just about every aspect of natural language (with the exception of short-distance dependences) are being violated by the pseudo-sentences.
These include lexical relationships, topic and discourse coherence, syntax and seman-
tics. One would expect that such glaring deficienci These include lexical relationships, topic and discourse coherence, syntax and semantics. One would expect that such glaring deficiencies in this simple model would be quickly remedied. Not so. We will now review these asp tics. One would expect that such glaring deficiencies in this simple model would be

Phil. Trans. R. Soc. Lond. A (2000)

AATHEMATICAL,
HYSICAL
k ENGINEERING

Ξ

PHILOSOPHICAL
TRANSACTIONS

*AATHEMATICAL,
'HYSICAL
k ENGINEERING
CENCES*

Downloaded from rsta.royalsocietypublishing.org

Incorporatinglinguisticstructure ¹³¹³

Table 2. *Trigram-generated pseudo-sentences*

(Average length pseudo-sentences generated by a trigram trained on the BN corpus.)

Verage length pseudo-sentences generated by a trigram trained on the BN corpus.
YOU CALL PORK MITCHELL IS THOSE THREE WIRE LUCK AFTER
ATTENDANT S_COMPETITIVENESS_AND_KNOWS_THAT </s> YOU CALL PORK MITCHELL IS THOSE THREE WIRE LUCK AFTENT ATTENDANT S. COMPETITIVENESS AND KNOWS THAT </s> ATTENDANT S. COMPETITIVENESS AND KNOWS THAT \langle /s \rangle
ARE YOU REFERRING TO IS EXTREMELY RISKY BECAUSE I'VE BEEN

TESTED WHOSE ONLY WITH A MAIN \langle /s $>$

THE FIRST BLACK EDUCATORS CATACOMBS DOWN ROMAN GABRIEL SLEEP IN A WAY TO KNOW IS PROPER \langle /s $>$

GABRIEL SLEEP IN A WAY TO KNOW IS PROPER \langle /s>MY QUESTION TO YOU THOSE PICTURES MAY STILL NOT IN ROMANIA AND I LOOKED UP CLEAN \langle /s> MY QUESTION TO YOU THOSE PICTURES MAY
ROMANIA AND I LOOKED UP CLEAN </s>

ROMANIA AND I LOOKED UP CLEAN </s>NOW ANCHORAGE LIFTED YOU WERE GOING TO TAKE THEIR CUE FROM ANCHORAGE LIFTED OFF EVERYTHING WILL WORK SITE VERDI NOW WERE GOING TO TAKE THEIR CUE FROM ANCH OFF EVERYTHING WILL WORK SITE VERDI \langle /s> (*b*) *Lexical relations*

To an *n*-gram, the vocabulary is a long list of indistinguishable categories. But of course, words in a language form complex and not fully understood lexical relations. To an *n*-gram, the vocabulary is a long list of indistinguishable categories. But of course, words in a language form complex and not fully understood lexical relations.
Surely TUESDAY is closer in some sense to WEDNESDA To an *n*-gram, the vocabulary is a long list of indistinguishable categories. But of course, words in a language form complex and not fully understood lexical relations. Surely TUESDAY is closer in some sense to WEDNESDA

Surely TUESDAY is closer in some sense to WEDNESDAY than to, say, CHAIR.
The simplest attempt to consider lexical relations concentrates on part-of-speech (POS) information. The POS-based *n*-gram (Jelinek 1989) comes in The simplest attempt to consider lexical relations concentrates on part-of-speech (POS) information. The POS-based n -gram (Jelinek 1989) comes in several varieties. For example, for a trigram, one could try

$$
Pr(w_i | w_{i-2}, w_{i-1}) = Pr(w_i | POS_i) \cdot Pr(POS_i | POS_{i-2}, POS_{i-1}),
$$
 (2.3)

 $Pr(w_i | w_{i-2}, w_{i-1}) = Pr(w_i | POS_i) \cdot Pr(POS_i | POS_{i-2}, POS_{i-1}),$ (2.3)
where POS_i is the POS class of w_i . The main motivation for such a model is to
reduce the number of parameters and hence the variance of the estimation. One where POS_i is the POS class of w_i . The main motivation for such a model is to reduce the number of parameters and, hence, the variance of the estimation. One practical problem is that in a language as polysemous as Engl where POS_i is the POS class of w_i . The main motivation for such a model is to reduce the number of parameters and, hence, the variance of the estimation. One practical problem is that in a language as polysemous as Engl reduce the number of parameters and, hence, the variance of the estimation. One
practical problem is that in a language as polysemous as English, the correct POS of
each word token is often hard to determine. State-of-thepractical problem is that in a language as polysemous as English, the correct POS of
each word token is often hard to determine. State-of-the-art POS taggers, boasting
95–97% accuracy under ideal conditions, can be helpful each word token is often hard to determine. State-of-the-art POS taggers, boasting 95–97% accuracy under ideal conditions, can be helpful. Alternatively, a hidden-
variable model can be used, in which all possible POSs are 95–97% accuracy under ideal conditions, can be helpful. Alternatively, a hiddenvariable model can be used, in which all possible POSs are considered simultaneously.
Nonetheless, these models are not usually very successful variable model can be used, in which all possible POSs are considered simultaneously.
Nonetheless, these models are not usually very successful, as measured by perplexity
improvement over the baseline word-based *n*-gram. Nonetheless, these models are not usually very successful, as measured by per
improvement over the baseline word-based n -gram. Apparently, what is a
linguistic distinction does not translate into a useful predictive dis

provement over the baseline word-based n -gram. Apparently, what is a useful guistic distinction does not translate into a useful predictive distinction.
An improvement over the POS-based model is to use a class-based mo linguistic distinction does not translate into a useful predictive distinction.
An improvement over the POS-based model is to use a class-based model, where
classes may get their origin in POS categories, but are further o An improvement over the POS-based model is to use a class-based model, where classes may get their origin in POS categories, but are further optimized over the data. Several algorithms have been suggested for automaticall classes may get their origin in POS categories, but are further optimized over the data. Several algorithms have been suggested for automatically clustering the vocabulary based on information-theoretic measures (e.g. Brow data. Several algorithms have been suggested for automatically clustering the vocabulary based on information-theoretic measures (e.g. Brown *et al.* 1991; Kneser & Ney 1993), in an either bottom-up or top-down fashion. I ulary based on information-theoretic measures (e.g. Brown *et al.* 1991; Kneser & Ney 1993), in an either bottom-up or top-down fashion. In some of these, the algorithm yields not just a partition into classes but rather (1993), in an either bottom-up or top-down fashion. In some of these, the algorithm yields not just a partition into classes but rather a word tree, namely a complete \bigcup (usually binary) hierarchy of word types. These yields not just a partition into classes but rather a word tree, namely a complete (usually binary) hierarchy of word types. These classes are then used by an *n*-gram similar to the one in equation (2.3). Yet another var (usually binary) hierarchy of word types. These classes are then used by an n -gram similar to the one in equation (2.3) . Yet another variation is to assume that each word type can belong to several different categorie similar to the on
word type can be
variable model.
Examples of w

O word type can belong to several different categories ('soft classes'), and use a hidden-

Examples of word classes derived by S. F. Chen (1998, unpublished work) using

such an algorithm are shown in table 3. Note that a variable model.
Examples of word classes derived by S. F. Chen (1998, unpublished work) using
such an algorithm are shown in table 3. Note that although most of the members
of a class seem appropriate some are not. Not sur Examples of word classes derived by S. F. Chen (1998, unpublished work) using
such an algorithm are shown in table 3. Note that although most of the members
of a class seem appropriate, some are not. Not surprisingly, the such an algorithm are shown in table 3. Note that although most of the members
of a class seem appropriate, some are not. Not surprisingly, the 'misfits' are often
rare word types, which only occurred a handful of times in of a class seem appropriate, some are not. Not surprisingly, the 'misfits' are often
rare word types, which only occurred a handful of times in the data on which the
clustering algorithm was run. Ironically, it is exactly *Phil. Trans. R. Soc. Lond.* A (2000)

**IYSICAL
ENGINEERING** ATHEMATICAL

OYAL

Ξ

1314 *Downloaded from rsta.royalsocietypublishing.or*
1314 *R. Rosenfeld* Downloaded from rsta.royalsocietypublishing.org

Table 3. *Automatically derived word classes*

(Word classes derived automatically from data; notice the 'misfits' are infrequent words.)

MY THY JESSICA'S SARAH'S KEVIN'S CONGESTIVE KAREN'S HEIDI'S THEN THEREFORE CONSEQUENTLY THIRDLY LASTLY BEHOLD FRO ABETTING

THEN THERE ORE CONSEQUENTET THRUST ENSTIT BEHOLD TRUNK
DOWN ASIDE ASHORE INS OVERBOARD IDLY... AFIRE ROUGHSHOD DOWN ASIDE ASHORE INS OVERBOARD IDLY
LET EXCUSE FORGIVE PARDON TICKLE

LET EXCUSE FORGIVE PARDON TICKLE
STATE CENSUS COMMONWEALTH PROVISIONAL FOOTHILLS

WASHINGTON LONDON MOSCOW PARIS TOKYO: : :ISLAMABAD EDGEWISE

DONE RESOLVED ACCOMPLISHED ACHIEVED FORGOTTEN SOLVED TOLERATED UNDERTAKEN NOTS FORESEEN

FOLERATED UNDERTAKEN NOTS FORESEEN
end of the vocabulary distribution, that stood to benefit the most from clustering.
This is true for all data-driven vocabulary-clustering algorithms: the more common end of the vocabulary distribution, that stood to benefit the most from clustering.
This is true for all data-driven vocabulary-clustering algorithms: the more common
the word is the more reliably it can be assigned to an end of the vocabulary distribution, that stood to benefit the most from clustering.
This is true for all data-driven vocabulary-clustering algorithms: the more common
the word is, the more reliably it can be assigned to an This is true for all data-driven vocabulary-clumble the word is, the more reliably it can be assigned.
Less it will benefit from such an assignment.
For this and other reasons, class-based n -g For this and other reasons, class-based n-gram models have only seen moderate
For this and other reasons, class-based n-gram models have only seen moderate
ccess. For any amount of training data, these models do not perfo

success it will benefit from such an assignment.
For this and other reasons, class-based n -gram models have only seen moderate
success. For any amount of training data, these models do not perform as well as
their word-For this and other reasons, class-based n -gram models have only seen moderate success. For any amount of training data, these models do not perform as well as their word-based counterparts. When the two are interpolated success. For any amount of training data, these models do n
their word-based counterparts. When the two are interpolate
improvement is usually achieved, but only for large corpora.
The only circumstance where lexical relat their word-based counterparts. When the two are interpolated together, a modest
improvement is usually achieved, but only for large corpora.
The only circumstance where lexical relations are exploited successfully for lan-

improvement is usually achieved, but only for large corpora.
The only circumstance where lexical relations are exploited successfully for language modelling is in very narrow discourse domains, where class-based *n*-gram The only circumstance where lexical relations are exploited successfully for language modelling is in very narrow discourse domains, where class-based *n*-gram models are used with hand-tailored classes. For example, in t guage modelling is in very narrow discourse domains, where class-based *n*-gram models are used with hand-tailored classes. For example, in the Airline Travel Information System (ATIS) domain (Price 1990), classes consisti els are used with hand-tailored classes. For example, in the Airline Travel Information System (ATIS) domain (Price 1990), classes consisting of city names, airline names, aircraft types, etc., proved very useful in the fa System (ATIS) domain (Price 1990), classes consisting of city names, airline names,

(*c*) *Syntactic structure*

Several attempts have been made to integrate theories of syntax into language modelling. We will mention three of them here.

(i) *Probabilistic context-free grammars*

Probabilistic context-free grammars
Context-free grammars (CFGs) are inaccurate as models of natural language, yet
n arguably serve as a first-order approximation. Probabilistic context-free gram-Context-free grammars (CFGs) are inaccurate as models of natural language, yet
can, arguably, serve as a first-order approximation. Probabilistic context-free gram-
mars (PCFGs) are CFGs with a probability distribution de Context-free grammars (CFGs) are inaccurate as models of natural language, yet
can, arguably, serve as a first-order approximation. Probabilistic context-free gram-
mars (PCFGs) are CFGs with a probability distribution def can, arguably, serve as a first-order approximation. Probabilistic context-free grammars (PCFGs) are CFGs with a probability distribution defined over all productions that share their left-hand side. To use PCFGs to model mars (PCFGs) are CFGs with a probability distribution defined over all productions
that share their left-hand side. To use PCFGs to model unconstrained language,
one must decide on both the CFG itself (set of non-terminals that share their left-hand side. To use PCFGs to model unconstrained language, \bullet one must decide on both the CFG itself (set of non-terminals and production rules). and the (usually context-free) production probabilities. To date, no CFG has been suggested that sufficiently covers unconstrained English. Given a large parsed and annotated corpus such as the Penn Treebank (Marcus *et al.* 1993), a CFG can be created to cover it, although its coverage of new, unseen annotated corpus such as the Penn Treebank (Marcus *et al.* 1993), a CFG can be created to cover it, although its coverage of new, unseen data will be more limited. Furthermore, given a CFG and annotated data, the 'inside created to cover it, although its coverage of new, unseen data will be more limited.

**YSICAL
INGINEERING THEMATICAL**

Ξ

HEMATICA

UAXO

PHILOSOPHICAL
TRANSACTIONS

OYAL

H

PHILOSOPHICAL
TRANSACTIONS

probabilities. However, the local optima found by the algorithm are unlikely to be as probabilities. However, the local optima found by the algorithm are unlikely to be as
good as the global optimum, which is computationally infeasible to find. Even if the
global optimum were to be found, it is likely that probabilities. However, the local optima found by the algorithm are unlikely to be as
good as the global optimum, which is computationally infeasible to find. Even if the
global optimum were to be found, it is likely that global optimum were to be found, it is likely that context-free production probabili-
ties do not have sufficient expressive power to capture the true distribution of parses.
For these reasons, no PCFGs have been suggested global optimum were to be found, it is likely that context-free production probabili-
ties do not have sufficient expressive power to capture the true distribution of parses.
For these reasons, no PCFGs have been suggested ties do not have sufficient expressive power to capture
For these reasons, no PCFGs have been suggested t
with the conventional *n*-gram, let alone surpass it.
An interesting attempt to combine *n*-grams and

with the conventional n -gram, let alone surpass it.
An interesting attempt to combine n -grams and PCFGs was reported by Miller (1995) . The CFG structure was formulated as a Markov random field (MRF) , and a family of additional constraints was imposed on transitions between successive (1995). The CFG structure was formulated as a Markov random field (MRF), and
a family of additional constraints was imposed on transitions between successive
words, effectively capturing bigram information. This fusing of a family of additional constraints was imposed on transitions between successive
words, effectively capturing bigram information. This fusing of CFG and bigrams
resulted in a model with size (number of parameters) comparab words, effectively capturing bigram information. This fusing of CFG and bigrams
resulted in a model with size (number of parameters) comparable with a bigram, yet
performance comparable with that of a trigram. However, no Equivalently in a model with size (number of parameters) comparable with a bigram, yet
performance comparable with that of a trigram. However, no improvement over the
state-of-the-art trigram has been reported.

(ii) *Probabilistic link grammar*

Link grammar is a lexicalized grammar formalism introduced by Sleator & Tem-
perley (1991), where a specific link grammar for English has also been constructed Link grammar is a lexicalized grammar formalism introduced by Sleator & Temperley (1991), where a specific link grammar for English has also been constructed
by hand, with encouraging coverage. In a specialized form of th Link grammar is a lexicalized grammar formalism introduced by Sleator & Temperley (1991), where a specific link grammar for English has also been constructed by hand, with encouraging coverage. In a specialized form of th perley (1991), where a specific link grammar for English has also been constructed
by hand, with encouraging coverage. In a specialized form of the grammar known as
'grammatical trigrams' (Lafferty *et al.* 1992), a word c by hand, with encouraging coverage. In a specialized form of the grammar known as

"grammatical trigrams" (Lafferty *et al.* 1992), a word can be predicted from any pair

of adjacent words that precede it in the sentence. 'grammatical trigrams' (Lafferty *et al.* 1992), a word can be predicted from any pair of adjacent words that precede it in the sentence. The choice of which such pair to use is encoded in the link grammar, which is train use is encoded in the link grammar, which is trained automatically from a corpus. use is encoded in the link grammar, which is trained automatically from a corpus.
Grammatical trigrams have achieved a modest yet consistent perplexity improvement
over the state-of-the-art trigram. Other promising forms Grammatical trigrams have achieved a modest yet consistent perplexity implement the state-of-the-art trigram. Other promising forms of a dependency were also attempted (Stolcke *et al.* 1997; Alshawi & Douglas, this issue) were also attempted (Stolcke *et al.* 1997; Alshawi & Douglas, this issue).
(iii) *Structured language model*

i) Structured language model
Recently, Chelba & Jelinek (1999) introduced a model that predicts the next word
sed on a set of linguistic equivalence classifications of the history. Given a history **IATHEMATICAL,
HYSICAL
< ENGINEERING
CIENCES** $\frac{1}{2}$ Becently, Chelba & Jelinek (1999) introduced a model that predicts the next word
based on a set of linguistic equivalence classifications of the history. Given a history,
a lexicalized parser proposes several po Recently, Chelba & Jelinek (1999) introduced a model that predicts the next word
based on a set of linguistic equivalence classifications of the history. Given a history,
a lexicalized parser proposes several possible equi based on a set of linguistic equivalence classifications of the history. Given a history, a lexicalized parser proposes several possible equivalence classifications, each with its own weight. The predictions from the vario own weight. The predictions from the various classifications are combined linearly.
The parser uses a natural probabilistic parametrization of a push-down automaton,
and an EM algorithm is used for training. Experiments on The parser uses a natural probabilistic parametrization of a push-down automaton, and an EM algorithm is used for training. Experiments on the Switchboard corpus (Godfrey *et al.* 1992) show modest improvements in both per The parser uses a natural probabilistic parametrization of a push-down automaton, and an EM algorithm is used fo
(Godfrey *et al.* 1992) show mode
rate over the baseline trigram. rate over the baseline trigram.
(*d*) *Topic and semantic coherence*

One of the most striking aspects of the pseudo-sentences in table 2 is their lack One of the most striking aspects of the pseudo-sentences in table 2 is their lack
of topic and semantic coherence. There is a strong sense in reading these sentences
that they are not *about* anything One of the most striking aspects
of topic and semantic coherence. That they are not *about* anything. So that they are not *about* anything.
 $\begin{bmatrix}\n\mathbf{a} \\
\mathbf{b}\n\end{bmatrix}$ (i) *Model interpolation*

The earliest attempts to capture topic coherence were through the use of interpo-The earliest attempts to capture topic coherence were through the use of interpo-
lated language models. Typically, the training data were partitioned into multiple
sets each containing documents about a particular topic The earliest attempts to capture topic coherence were through the use of interpo-
lated language models. Typically, the training data were partitioned into multiple
sets, each containing documents about a particular topic *Phil. Trans. R. Soc. Lond.* A (2000)

1316 $R. Rosenfeld$
set was used to create a separate topic-specific language model $P_t(w | h)$, and the
various models were interpolated together at the word level. set was used to create a separate topic-specific language mod
various models were interpolated together at the word level, various models were interpolated together at the word level,

$$
P(w | h) = \sum_{t} \lambda_t \cdot P_t(w | h), \qquad (2.4)
$$

where the interpolation weights $\{\lambda_1, \lambda_2, ...\}$ varied based on the expected topic of
the test data, and were generally determined from held-out data. where the interpolation weights $\{\lambda_1, \lambda_2, \dots\}$ varied based on the ϵ the test data, and were generally determined from held-out data.
There are many variations on this general approach. The training here the interpolation weights $\{\lambda_1, \lambda_2, \dots\}$ varied based on the expected topic of e test data, and were generally determined from held-out data.
There are many variations on this general approach. The training data m

the test data, and were generally determined from held-out data.
There are many variations on this general approach. The training data may be provided already classified into topics (e.g. Seymore & Rosenfeld 1997), or a c There are many variations on this general approach. The training data may be provided already classified into topics (e.g. Seymore $\&$ Rosenfeld 1997), or a clustering algorithm may need to be run to automatically derive vided already classified into topics (e.g. Seymore & Rosenfeld 1997), or a clustering algorithm may need to be run to automatically derive such classification (e.g. Iyer & Ostendorf 1999). The topic classes themselves can algorithm may need to be run to automatically derive such classification (e.g. Iyer & Ostendorf 1999). The topic classes themselves can be hard, soft (i.e. allow overlaps), or can even be arranged to form a hierarchy (Sey Ostendorf 1999). The topic classes themselves can be hard, soft (i.e. allow overlaps), or can even be arranged to form a hierarchy (Seymore & Rosenfeld 1997). Finally, interpolation can take place at the word level, as in or can even be a
interpolation can
sentence level,

$$
P(s) = \sum_{t} \lambda_t \cdot P_t(s) = \sum_{t} \lambda_t \cdot \prod_{i} P_t(w_i \mid h_i), \qquad (2.5)
$$

or at both (Iyer & Ostendorf 1999). Generally speaking, topic interpolation results in
moderate vet consistent reductions in perplexity, and often also in speech-recognition or at both (Iyer & Ostendorf 1999). Generally speaking, topic interpolation results in moderate yet consistent reductions in perplexity, and often also in speech-recognition error rates. or at both (I₎
moderate yet
error rates.
However. i moderate yet consistent reductions in perplexity, and often also in speech-recognition
error rates.
However, interpolation is seriously deficient as a method for modelling topic coher-

ence. This is because it fails to separate those aspects of language that vary from
topic to topic from those that are invariant across all topics. As a result, the limited
topic to topic from those that are invariant acro However, interpolation is seriously deficient as a method for modelling topic coher-
ence. This is because it fails to separate those aspects of language that vary from
topic to topic from those that are invariant across a topic to topic from those that are invariant across all topics. As a result, the limited be pulled in for more robust estimation, resulting in a dilution in the topicality of the interpolated model. amount of training data in each topic means that the out-of-topic training data must

(ii) *Cache*

Cache
Another attempt to capture topic coherence and word correlations was through
e use of an *n*-gram cache (Kuhn & De Mori 1990). Caches are easy to implement Another attempt to capture topic coherence and word correlations was through
the use of an n-gram cache (Kuhn & De Mori 1990). Caches are easy to implement,
and capture word auto-correlations, which are a very pronounced Another attempt to capture topic coherence and word correlations was through
the use of an *n*-gram cache (Kuhn & De Mori 1990). Caches are easy to implement,
and capture word auto-correlations, which are a very pronounce the use of an *n*-gram cache (Kuhn & De Mori 1990). Caches are easy to implement, and capture word auto-correlations, which are a very pronounced phenomenon across sentences. Both Kuhn & De Mori (1990) and Jelinek *et al.* and capture word auto-correlations, which are a very pronounced phenomenon across
sentences. Both Kuhn & De Mori (1990) and Jelinek *et al.* (1991) report improve-
ments in perplexity over the baseline trigram, and the la sentences. Both Kuhn & De Mori (1990) and Jelinek *et al.* (1991) report improve-
ments in perplexity over the baseline trigram, and the latter group also reports a
modest reduction in word-recognition error rate. Since t ments in perplexity over the baseline trigram, and the latter group also reports a modest reduction in word-recognition error rate. Since then, caches have been implemented in many systems, with similar results, and have n modest reduction in word-recognition
mented in many systems, with simi
'baseline' in language modelling.† (iii) *Word triggers*

 $\begin{pmatrix} 0 & \text{if } 0 \\ \text{if } 0 & \text{if } 0 \end{pmatrix}$ M generalization of the cache idea to correlations between different words led to work on *word triggers* (Rosenfeld 1996; Beeferman *et al*. 1997). In principle, cor-A generalization of the cache idea to correlations between different words led to
work on *word triggers* (Rosenfeld 1996; Beeferman *et al.* 1997). In principle, cor-
relations between any pair of words or phrases can be work on *word triggers* (Rosenfeld 1996; Beeferman *et al.* 1997). In principle, correlations between any pair of words or phrases can be captured and modelled. In practice, Rosenfeld (1996) showed that linear interpolati relations between any pair of words or phrases can be captured and modelled. In practice, Rosenfeld (1996) showed that linear interpolation of the trigger component is suboptimal, and that an exponential model, trained usi is suboptimal, and that an exponential model, trained using the maximum-entropy
† We did not use a cache in generating the sentences in table 2 because these sentences are evaluated

in isolation, whereas the autocorrelations a cache is designed to capture are predominantly cross-sentence effects.

Phil. Trans. R. Soc. Lond. A (2000)

AATHEMATICAL,
HYSICAL
k ENGINEERING

E

Incorporatinglinguisticstructure ¹³¹⁷ Incorporating linguistic structure 1317
principle, is superior. Unfortunately, the computational requirements of training such
a model grow supra-linearly with the number of independently modelled word trig-Downloaded from rsta.royalsocietypublishing.org

INEERING
IES **ATHEMATICAL**

PHILOSOPHICAL
TRANSACTIONS

EERING **ATHEMATICAL**

UNXO

THE

PHILOSOPHICAL
TRANSACTIONS

(iv) *Dimensionality reduction*

An improvement over modelling individual word correlations can be achieved by using singular value decomposition (SVD) to reduce the dimensionality of the topic space. In Bellegarda (1998), a matrix of word-document occurrences is reduced to a An improvement over modelling individual word correlations can be achieved by using singular value decomposition (SVD) to reduce the dimensionality of the topic space. In Bellegarda (1998), a matrix of word-document occur using singular value decomposition (SVD) to reduce the dimensionality of the topic
space. In Bellegarda (1998), a matrix of word-document occurrences is reduced to a
relatively small size (100 \times 100) via SVD. The resul space. In Bellegarda (1998), a matrix of word-document occurrences is reduced to a
relatively small size (100 \times 100) via SVD. The resulting matrix succinctly captures
the most salient correlations between groups of wor relatively small size (100×100) via SVD. The resulting matrix succinctly captures
the most salient correlations between groups of words on the one hand and clusters
of documents on the other. The SVD process also prov the most salient correlations between groups of words on the one hand and clusters
of documents on the other. The SVD process also provides the necessary projections
from document-space and word-space into the new, combine of documents on the other. The SVD process also provides the necessary projections
from document-space and word-space into the new, combined space. As a result,
any new document or partial document can be projected into th from document-space and word-space into the new, combined space. As a result, any new document or partial document can be projected into the combined space, effectively being classified as a combination of the 100 underly any new document or partial document can be projected into the combined space, effectively being classified as a combination of the 100 underlying semantic dimensions. When combining SVD decomposition with an n -gram, $\frac{1}{0}$ in perplexity are reported, as well as in speech-recognition errors (Bellegarda 2000).

principle, is superior. Unfortunately, the computational requirements of training such
a model grow supra-linearly with the number of independently modelled word trig-
ger pairs, and are prohibitive even for a moderate num principle, is superior. Unfortunately, the computational requirements of training such
a model grow supra-linearly with the number of independently modelled word trig-
ger pairs, and are prohibitive even for a moderate num a model grow supra-linearly with the number of independently modelled word trigger pairs, and are prohibitive even for a moderate number of such pairs. Although such a model achieves significant perplexity reduction over t such a model achieves significant perplexity reduction over the baseline trigram, the computational difficulties render it impractical in most cases of interest.

3. A general framework for integrating linguistic structure

3. A general framework for integrating linguistic structure
The modelling attempts described in the previous section suffer from two major
deficiencies. First, the statistical methodology in these attempts varied greatly. The modelling attempts described in the previous section suffer from two major
deficiencies. First, the statistical methodology in these attempts varied greatly. Each
such model was aimed at a specific linguistic phenomeno The modelling attempts described in the previous section suffer from two major deficiencies. First, the statistical methodology in these attempts varied greatly. Each such model was aimed at a specific linguistic phenomeno deficiencies. First, the statistical methodology in these attempts varied greatly. Each
such model was aimed at a specific linguistic phenomenon, which, in turn, affected
the choice of model structure, parameter family, tr such model was aimed at a specific linguistic phenomenon, which, in turn, affected
the choice of model structure, parameter family, training algorithms, etc. In addition,
a new method had to be found for combining the new the choice of model structure, parameter family, training algorithms, etc. In addition,
a new method had to be found for combining the new model component with the
existing n-gram baseline. If a new linguistic knowledge so a new method had to be found for combining the new model component with the existing n -gram baseline. If a new linguistic knowledge source were to suggest itself, a new modelling methodology would have to be developed a existing *n*-gram baseline. If a new linguistic knowledge sou
a new modelling methodology would have to be develope
practical estimation issues would have to be worked out.
Second, virtually all the models described above new modelling methodology would have to be developed and tested, and many
actical estimation issues would have to be worked out.
Second, virtually all the models described above estimate the probability of a sen-
nce s by

practical estimation issues would have to be worked out.
Second, virtually all the models described above estimate the probability of a sentence s by using the chain rule, as in equation (2.1), to break it into a product Second, virtually all the models described above estimate the probability of a sentence s by using the chain rule, as in equation (2.1), to break it into a product of conditional probabilities (typically $P(w | h)$). While t tence s by using the chain rule, as in equation (2.1), to break it into a product of conditional probabilities (typically $P(w | h)$). While this practice is understandable from a historical perspective (*n*-gram modelling c ditional probabilities (typically $P(w | h)$). While this practice is understandable from
a historical perspective (*n*-gram modelling cannot be done on whole sentences), it is
not desirable for capturing linguistic phenomen a historical perspective (*n*-gram modelling cannot be done on whole sentences), it is
not desirable for capturing linguistic phenomena. Linguistic aspects of sentences—
such as their grammar, syntax, semantics or pragmat not desirable for capturing linguistic phenomena. Linguistic aspects of sentences—
such as their grammar, syntax, semantics or pragmatics—are impossible or at best
awkward to think about, let alone encode, in a conditional such as their grammar, syntax, semantics or pragmatics—are impossible or at best
awkward to think about, let alone encode, in a conditional framework. Also, external
influences on the sentence (e.g. the effect of preceding awkward to think about, let alone encode, in a conditional framework. Also, external
influences on the sentence (e.g. the effect of preceding utterances, or dialogue-level
variables) are equally hard to encode, and factori influences on the sentence (e.g. the effect of preceding utterances, or dialogue-level
variables) are equally hard to encode, and factoring them into the prediction of every
word in the current sentence causes small but sy variables) are equally hard to end
word in the current sentence ca
estimation to be compounded.
We have recently introduced a word in the current sentence causes small but systematic biases in the probability estimation to be compounded.
We have recently introduced a new language-modelling framework that addresses

estimation to be compounded.
We have recently introduced a new language-modelling framework that addresses
these two deficiencies (Rosenfeld 1997). The exponential model we use directly mod-
els the probability of an entir We have recently introduced a new language-modelling framework that addresses
these two deficiencies (Rosenfeld 1997). The exponential model we use directly mod-
els the probability of an entire sentence or utterance. By these two deficiencies (Rosenfeld 1997). The exponential model we use directly models the probability of an entire sentence or utterance. By avoiding the chain rule, the model treats each sentence or utterance as a 'bag of els the probability of an entire sentence or utterance. By avoiding the chain rule, the model treats each sentence or utterance as a 'bag of features',† where features are arbitrary computable properties of the sentence. F the model treats each sentence or utterance as a 'bag of features',† where features are arbitrary computable properties of the sentence. Furthermore, the unified structure of the model means that any linguistic theory can O ture of the model means that any linguistic theory can be incorporated without any
 \uparrow Not to be confused with a *bag of words*: features may take account of sequentiality, if so desired.

¹³¹⁸ *R. Rosenfeld* Downloaded from rsta.royalsocietypublishing.org

1318 $R. Rosenfeld$
change to the model itself. This solves the two problems mentioned above. In this
section we describe the model and review the various features it has been used with change to the model itself. This solves the two problems mentioned above. In this section we describe the model and review the various features it has been used with so far. section we describe the model and review the various features it has been used with so far.

(*a*) *A whole-sentence exponential model*

(a) A whole-sentence exponential model
A whole-sentence exponential language model has the form

onential language model has the form
\n
$$
P(s) = \frac{1}{Z} \cdot P_0(s) \cdot \exp\left[\sum_i \lambda_i f_i(s)\right],
$$
\n(3.1)

 $F(s) = \frac{1}{Z} \cdot F_0(s) \cdot \exp\left[\frac{1}{Z} \cdot \frac{\lambda_i f_i(s)}{s}\right],$ (3.1)
where the λ_i are the parameters of the model, Z is a universal normalization constant
that depends only on the λ_i and the $f_i(s)$ are arbitrary computable prope where the λ_i are the parameters of the model, Z is a universal normalization constant
that depends only on the λ_i , and the $f_i(s)$ are arbitrary computable properties, or
features, of the sentence s. The distribution where the λ_i are the parameters of the model, Z is a universal normalization constant
that depends only on the λ_i , and the $f_i(s)$ are arbitrary computable properties, or
features, of the sentence s. The distributi that depends only on the λ_i , and the $f_i(s)$ are arbitrary computable properties, or *features*, of the sentence s. The distribution $P_0(s)$ is an arbitrary probability distribution. It can be thought of as the starting **improvements**. Often, P₀(s) is an arbitrary probability distri-
 improvements. Often, P₀(s) will be simply derived from the baseline trigram.
 improvements. Often, P₀(s) will be simply derived from the baselin

bution. It can be thought of as the starting point, or baseline, for further modelling
improvements. Often, $P_0(s)$ will be simply derived from the baseline trigram.
The features $\{f_i(s)\}$ are selected by the modeller to grams, longer-distance dependences, or simple global sentence properties, to more data they consider appropriate or profitable. These can vary from conventional *n*-
grams, longer-distance dependences, or simple global sentence properties, to more
complex functions based on POS tagging, parsing, or othe grams, longer-distance dependences, or simple global senter complex functions based on POS tagging, parsing, or other ty (person and number agreement, semantic coherence, etc.).
For each feature $f_i(s)$ its expectation und mplex functions based on POS tagging, parsing, or other types of linguistic analysis
erson and number agreement, semantic coherence, etc.).
For each feature $f_i(s)$, its expectation under $P(s)$ is constrained to a specific

 K_i :

$$
E_P f_i = K_i. \tag{3.2}
$$

These target values are typically set to the expectation of that feature under the These target values are typically set to the expectation of that feature under the
empirical distribution \tilde{P} of the training corpus $T = \{s_1, \ldots, s_N\}$ (for binary features,
this is simply the prevalence of that feat These target values are typically set to the expectation of that feature under the empirical distribution \tilde{P} of the training corpus $T = \{s_1, \ldots, s_N\}$ (for binary features, this is simply the prevalence of that feat becomes

$$
\sum_{s} P(s) \cdot f_i(s) = E_{\tilde{P}} f_i \equiv \frac{1}{N} \sum_{j=1}^{N} f_i(s_j).
$$
 (3.3)

If the constraints (3.2) are consistent, there exists a unique solution $\{\lambda_i\}$ within the exponential family (3.1) that satisfies them. Among all (not necessarily exponential) If the constraints (3.2) are consistent, there exists a unique solution $\{\lambda_i\}$ within the exponential family (3.1) that satisfies them. Among all (not necessarily exponential) solutions to equation (3.2), the exponentia exponential family (3.1) that satisfies them. Among all (not necessarily exponential) solutions to equation (3.2), the exponential solution is the one closest to the baseline $P_0(s)$ (in the Kullback–Liebler sense), and i solutions to equation (3.2), the exponential solution is the one closest to the baseline $P_0(s)$ (in the Kullback–Liebler sense), and is thus called the minimum-divergence or minimum-discrimination-information (MDI) solut $P_0(s)$ (in the Kullback–Liebler sense), and is thus called the minimum-divergence
or minimum-discrimination-information (MDI) solution. If the baseline $P(s)$ is flat
(uniform), this becomes the maximum-entropy (ME) solut or minimum-discrimination-information (MDI) solution. If the baseline $P(s)$ is flat
(uniform), this becomes the maximum-entropy (ME) solution. Furthermore, if the
feature target values K_i are the empirical expectations (uniform), this becomes the maximum-entropy (ME) solution. Furthermore, if the feature target values K_i are the empirical expectations over some training corpus (as in equation (3.3)), the MDI or ME solution is also the feature target values K_i are the empirical expectations over some training corpus (as
in equation (3.3)), the MDI or ME solution is also the maximum-likelihood solution
of the exponential family. For more information se O of the exponential family. For more information see Jaynes (1957), Berger *et al.*
 \bigcirc (1996) and Rosenfeld (1996).
 \bigcirc It is instructive to compare this model with the conditional exponential model,

(1996) and Rosenfeld (1996).
It is instructive to compare this model with the conditional exponential model,
which has seen considerable success recently in language modelling (Della Pietra *et*
 al 1992; Lau *et al* 1993 It is instructive to compare this model with the conditional exponential model, which has seen considerable success recently in language modelling (Della Pietra *et al.* 1992; Lau *et al.* 1993; Berger *et al.* 1996; Rosen which has seen cor
al. 1992; Lau et al.
has the form

$$
P(w | h) = \frac{1}{Z(h)} \cdot P_0(w | h) \cdot \exp\left[\sum_i \lambda_i f_i(h, w)\right],
$$
\n(3.4)

Phil. Trans. R. Soc. Lond. A (2000)

HH

PHILOSOPHICAL
TRANSACTIONS

**I ATHEMATICAL,
HYSICAL**
CENGINEERING

HL

Incorporating linguistic structure 1319
where the features are functions of a specific word-history pair, and so is the baseline
 P_0 . More importantly, Z is no longer a true constant; it depends on h and, thus, must where the features are functions of a specific word–history pair, and so is the baseline P_0 . More importantly, Z is no longer a true constant: it depends on h and, thus, must be recomputed for each word in each sentenc where the features are functions of a specific word-history pair, and so is the baseline P_0 . More importantly, Z is no longer a true constant: it depends on h and, thus, must be recomputed for each word in each sentenc P_0 . More importantly, Z is no longer a true constant: it depends on h and, thus, must
be recomputed for each word in each sentence. The main drawbacks of the conditional
model are the severe computational bottleneck of be recomputed for each word in each sentence. The main drawbacks of the conditional model are the severe computational bottleneck of training (especially of computing $Z(h)$), and the difficulty in modelling whole-sentence phenomena.
(b) *Training the model*

The MDI or ME solution can be found by an iterative procedure such as the The MDI or ME solution can be found by an iterative procedure such as the generalized-iterative-scaling (GIS) algorithm (Darroch & Ratcliff 1972). GIS starts with arbitrary λ_i . At each iteration, the algorithm improves The MDI or ME solution can be found by an iterative procedure such as the generalized-iterative-scaling (GIS) algorithm (Darroch & Ratcliff 1972). GIS starts with arbitrary λ_i . At each iteration, the algorithm improves with arbitrary λ_i . At each iteration, the algorithm improves the $\{\lambda_i\}$ values by comparing the expectation of each feature under the current P with the target value, and modifying the associated λ . In particular, Comparing the expectation of each feature under the current P with the target value,

$$
\lambda_i \qquad \lambda_i + F_i \log \frac{E_{\tilde{P}}[f_i]}{E_P[f_i]},\tag{3.5}
$$

 λ_i $\lambda_i + F_i \log \frac{E_f}{E_f}$
where F_i is a parameter affecting the step size.

(i) *Sampling*

Sampling
In training a whole-sentence maximum-entropy model, computing the expectations

$$
E_P[f_i] = \sum_s P(s) \cdot f_i(s)
$$

requires a summation over all possible sentences ^s, clearly an infeasible task. Instead, requires a summation over all possible sentences s, clearly an infeasible task. Instead, we estimate $E_P[f_i]$ by sampling from the distribution $P(s)$ and using the sample expectation of f. Sampling from an exponential dist requires a summation over all possible sentences s, clearly an infeasible task. Instead, we estimate $E_P[f_i]$ by sampling from the distribution $P(s)$ and using the sample expectation of f_i . Sampling from an exponential d we estimate $E_P[f_i]$ by sampling from the distribution $P(s)$ and using the sample
expectation of f_i . Sampling from an exponential distribution is a non-trivial task,
and is the subject of intense research by statistician expectation of f_i . Sampling from an exponential distribution is a non-trivial task, and is the subject of intense research by statisticians, physicists and others. Sampling of sentences from an exponential distribution and is the subject of intense research by statisticians, physicists and others. Sampling of sentences from an exponential distribution poses additional challenges, and is discussed in Chen & Rosenfeld (1999). Efficient sa training. discussed in Chen & Rosenfeld (1999). Efficient sampling is crucial to successful training.
It is equally infeasible to compute the normalization constant

$$
Z = \sum_{s} p_0(s) \cdot \exp\biggl(\sum_{i} \lambda_i f_i(s)\biggr).
$$

 $Z = \sum_{s} p_0(s) \cdot \exp\left(\sum_{i} \lambda_i f_i(s)\right)$.
Fortunately, this is not necessary for training, since sampling can be done without
knowing Z. Using the model as part of a classifier (e.g. a speech recognizer) does not Fortunately, this is not necessary for training, since sampling can be done without
knowing Z. Using the model as part of a classifier (e.g. a speech recognizer) does not
require knowledge of Z either, because the relativ Fortunately, this is not necessary for training, since sampling can be done without knowing Z . Using the model as part of a classifier (e.g. a speech recognizer) does not require knowledge of Z either, because the rel knowing Z . Using the model as part of a classifier (e.g. a speech recognizer) does not require knowledge of Z either, because the relative ranking of the different hypotheses is not changed by a single, universal, con require knowledge of Z either, because is not changed by a single, universeponential models.
Even though the exact value of not changed by a single, universal, constant. Notice that this is not the case for noticional exponential models.
Even though the exact value of Z is not really needed, at times it may be desirable
approximate it for ex

conditional exponential models.
Even though the exact value of Z is not really needed, at times it may be desirable
to approximate it, for example for perplexity calculation. This can be done to any
desired accuracy by ge Even though the exact value of Z is not really needed, at times it may be desirable
to approximate it, for example for perplexity calculation. This can be done to any
desired accuracy by generating a large sample from $P(s$ to approximate it, for example for perplexity calculation. This can be done to any desired accuracy by generating a large sample from $P(s)$, observing the frequency of one or more sentences that occur more than, say, 50 t desired accuracy by generating a large sample from $P(s)$, observing the frequency
of one or more sentences that occur more than, say, 50 times, and making use of
equation (3.1). For situations where no such sentences exis of one or more sentences that occur more
equation (3.1). For situations where no s
more efficient estimator, one could use

$$
\hat{Z} = \frac{1}{\|T_0\|} \sum_{s \in T_0} \left[\exp\left(\sum_i \lambda_i f_i(s)\right) \right],\tag{3.6}
$$

Phil. Trans. R. Soc. Lond. A (2000)

IATHEMATICAL,
HYSICAL
ENGINEERING
CIENCES

Ξ

Downloaded from rsta.royalsocietypublishing.or
1320 *R. Rosenfeld* Downloaded from rsta.royalsocietypublishing.org

1320 $R.$ *Rosenfeld*
where T_0 is a sample of sentences generated from P_0 . For more details, see Zhu *et* $\frac{d^2}{dx^2}$ where T_0 is where T_0 is $al.$ (1999).

(*c*) *Feature selection*

Once the general framework and training procedure have been worked out, attention can be concentrated on the art of modelling language. The goal is to choose Once the general framework and training procedure have been worked out, attention can be concentrated on the art of modelling language. The goal is to choose features $f_i(s)$ that capture aspects of language that are not c tion can be concentrated on the art of modelling language. The goal is to choose features $f_i(s)$ that capture aspects of language that are not captured (or inadequately captured) by the current baseline-modelling techniqu features $f_i(s)$ that capture aspects of language that are not captured (c
quately captured) by the current baseline-modelling technique. To this end,
been using the following methodology for feature discovery and selectio

quately captured) by the current baseline-modelling technique. To this end, we have
been using the following methodology for feature discovery and selection.
Given a corpus T of natural language sentences[†] with empirica Given a corpus T of natural language sentences† with empirical distribution P, presumably representative of the unknown target distribution P, we use it to train our best baseline model P_0 . Next, we use P_0 to genera presumably representative of the unknown target distribution P , we use it to train
our best baseline model P_0 . Next, we use P_0 to generate a corpus T_0 of 'pseudo-
sentences', like those in table 2. We then comp our best baseline model P_0 . Next, we use P_0 to generate a corpus T_0 of 'pseudo-
sentences', like those in table 2. We then compare T_0 with T (or some other dataset
from the same distribution P). We look fo sentences', like those in table 2. We then compare T_0 with T (or some other dataset
from the same distribution P). We look for systematic differences between the two
corpora. Any such difference we discover points from the same distribution P). We look for systematic differences between the two
corpora. Any such difference we discover points to a deficiency in the way P_0 models
the unknown target distribution P. Any such deficie corpora. Any such difference we discover points to a deficiency in the way P_0 models
the unknown target distribution P. Any such deficiency can now be readily fixed,
by defining an appropriate feature $f(s)$ (or set of the unknown target distribution P . Any such deficiency can now be readily fixed, by defining an appropriate feature $f(s)$ (or set of features) which have different expectations under P and P_0 (as evidenced by their by defining an appropriate feature $f(s)$ (or set of feature expectations under P and P_0 (as evidenced by their respections then added, resulting in a new model: The new feature is then added, resulting in a new model:

$$
P_1(s) = \frac{1}{Z} P_0(s) \cdot \exp^{\lambda f(s)}.
$$
 (3.7)

Once P_1 is trained, the appropriate constraint (equation (3.3)) guarantees that it \angle
Once P_1 is trained, the appropriate constraint (equation (3.3)) guarantees that it consistently captures the new feature, and the previously observed difference between our model and the target distribution has be Once P_1 is trained, the appropriate constraint (equation (
consistently captures the new feature, and the previously obse
our model and the target distribution has been eliminated.
The process can now be repeated by ge model and the target distribution has been eliminated.
The process can now be repeated by generating a corpus T_1 of 'pseudo-sentences'
om the improved model P_1 and comparing it with the original corpus T looking

our model and the target distribution has been eliminated.
The process can now be repeated by generating a corpus T_1 of 'pseudo-sentences'
from the improved model P_1 , and comparing it with the original corpus T , l The process can now be repeated by generating a corpus T_1 of 'pseudo-sentences'
from the improved model P_1 , and comparing it with the original corpus T , looking
for new differences. The latter will be captured wit from the improved model P_1 , and comparing it with the original corpus T , lookifor new differences. The latter will be captured with new features, and so on. practice, many features (or even sets of features) are adde The suppose that the captured with new features, and so on. In actice, many features (or even sets of features) are added at each iteration.[†] As an example, \P suppose we observe that the trigram-generated T_0 sentenc

practice, many features (or even sets of features) are added at each iteration.[†]
As an example,¶ suppose we observe that the trigram-generated T_0 sentences
are slightly shorter on average (as measured by number of wo As an example, \P suppose we observe that the are slightly shorter on average (as measured by counterparts. We then define the simple feature are slightly shorter on average (as measured by number of words) than their T
counterparts. We then define the simple feature
 $f_{\text{length}}(s) = \text{number of words in } s,$ (3.8)
and observe that $E_{P_0}[f_{\text{length}}] \neq E_{\tilde{P}}[f_{\text{length}}]$. But once

 $f_{\text{length}}(s) = \text{number of words in } s,$ (3.8)
and observe that $E_{P_0}[f_{\text{length}}] \neq E_{\tilde{P}}[f_{\text{length}}]$. But once the new feature is incorporated,
we are assured that $E_{P_0}[f_{\text{length}}] = E_{\tilde{P}}[f_{\text{length}}]$ $y_{\text{length}}(s) = \text{number of}$
and observe that $E_{P_0}[f_{\text{length}}] \neq E_{\tilde{P}}[f_{\text{length}}]$. But α
we are assured that $E_{P_1}[f_{\text{length}}] = E_{\tilde{P}}[f_{\text{length}}]$. $\begin{aligned} \n\text{gth} &= E_{\tilde{P}}[\tilde{f}_{\text{length}}]. \n\end{aligned}$ (*d*) The search for features

(d) The search for features
In Chen & Rosenfeld (1999), we searched for *n*-gram-style features that showed sig-
ficant discrepancy between P and P₀. These included 4-grams and 5-grams (which In Chen & Rosenfeld (1999), we searched for *n*-gram-style features that showed significant discrepancy between P and P_0 . These included 4-grams and 5-grams (which

Phil. Trans. R. Soc. Lond. A (2000)

PHILOSOPHICAL
TRANSACTIONS

EERING

icant discrepancy between P and P_0 . These included 4-grams and 5-grams (which
† Or, more generally, utterances. The model is equally suitable for direct estimation of any spoken
erance, whether or not it conforms to [†] Or, more generally, utterances. The model is equally suitable for direct utterance, whether or not it conforms to conventional linguistic boundaries.

[†] A process of iteratively incorporating the most information-bea \dagger Or, more generally, utterances. The model is equally suitable for direct estimation of any spoken
erance, whether or not it conforms to conventional linguistic boundaries.
 \ddagger A process of iteratively incorporating

into an exponential model was described in Della Pietra *et al.* (1997). The emphasis in our methodology, into an exponential model was described in Della Pietra *et al.* (1997). The emphasis in our methodology, though is \dagger A process of iteratively incorporating the most information-bearing feature in a given candidate set
into an exponential model was described in Della Pietra *et al.* (1997). The emphasis in our methodology,
though, i into an exponential model was described in Della Pietra *et al.* (1997). The *c* though, is on the manual inspection of two corpora and the linguistic ansearching for and evaluating families of linguistically motivated fe bugh, is on the manual inspection of two corpora and the linguistic analysis and 'detective work' of

arching for and evaluating families of linguistically motivated features.
 \blacksquare A true one, it turns out: properly sm

 \blacksquare A true one, it turns out: properly smoothed trigram models often do not accurately capture unigram marginals such as $Pr(<$ /s >), the end-of-sentence probability.

Incorporatinglinguisticstructure ¹³²¹ Downloaded from rsta.royalsocietypublishing.org

Incorporating linguistic structure 1321
were outside the range of the baseline P_0 trigram), class *n*-grams, and distance (non-
contiguous) *n*-grams. All such features were ranked by a χ^2 significance test. Over were outside the range of the baseline P_0 trigram), class *n*-grams
contiguous) *n*-grams. All such features were ranked by a χ^2 sign-
50,000 of these features were found to have a significance leve EERING ATHEMATICAL were outside the range of the baseline P_0 trigram), class *n*-grams, and distance (non-
contiguous) *n*-grams. All such features were ranked by a χ^2 significance test. Over
50 000 of these features were found to ha 50 000 of these features were found to have a significance level of $\chi^2 > 15$. When
incorporated into the language model, they resulted in a small improvement in recog-
intion accuracy (perplexity was not computed). Alth 50 000 of these features were found to have a significance level of $\chi^2 > 15$. When
incorporated into the language model, they resulted in a small improvement in recog-
nition accuracy (perplexity was not computed). Alth

incorporated into the language model, they resulted in a small improvement in recognition accuracy (perplexity was not computed). Although these n -gram features do not improve the linguistic plausibility of the model, t not improve the linguistic plausibility of the model, they served to verify and demon-
strate our methodology.

In Zhu *et al*. (1999), we used a shallow parser to map utterances from the Switchboard corpus into a flat list of variable length *constituents*. Features were then defined
in terms of constituent sequences, constituent sets and constituent trigrams. Some In Zhu *et al.* (1999), we used a shallow parser to map utterances from the Switchboard corpus into a flat list of variable length *constituents*. Features were then defined in terms of constituent sequences, constituent s board corpus into a flat list of variable length *constituents*. Features were then defined
in terms of constituent sequences, constituent sets and constituent trigrams. Some
7000 such features were found to be statistical in terms of constituent sequences, constituent sets and constituent trigrams. Some 7000 such features were found to be statistically significant and added to the model.
The perplexity of the new model was slightly lower th 7000 such features were found to be statistically significant and added to the model.
The perplexity of the new model was slightly lower than that of the baseline, and
recognition accuracy was also slightly improved. Furth The perplexity of the new model was slightly lower than the recognition accuracy was also slightly improved. Further and potential of these features was limited due to their rarity.
We have subsequently refocused our atten cognition accuracy was also slightly improved. Further analysis suggested that the tential of these features was limited due to their rarity.
We have subsequently refocused our attention on finding a small number of much o

potential of these features was limited due to their rarity.
We have subsequently refocused our attention on finding a small number of much
more common features. For example, among the most glaring differences between
true We have subsequently refocused our attention on finding a small number of much
more common features. For example, among the most glaring differences between
true natural language and trigram-generated sentences is the lack more common features. For example, among the most glaring differences between
true natural language and trigram-generated sentences is the lack of semantic and
topic coherence in the latter. We have been working on modelli true natural language and trigram-generated sentences is the lack of semantic and
topic coherence in the latter. We have been working on modelling such coherence
within this framework. As building blocks for the 'semantic topic coherence in the latter. We have been working on modelling such coherence within this framework. As building blocks for the 'semantic coherence' feature, we use measures of association in 2×2 contingency tables within this framework. As building blocks for the 'semantic coherence' feature, we use measures of association in 2×2 contingency tables based on pairs of content words in the same sentence. For more details, see Rosenfeld *et al.* (1999).
 $\bf{4.}$ Discussion

Why has the language modelling community failed thus far to integrate formal lin-Why has the language modelling community failed thus far to integrate formal linguistic theories into a statistical framework? Why do current practical language models lack any resemblance to even a rudimentary linguistic Why has the language modelling community failed thus far to integrate formal linguistic theories into a statistical framework? Why do current practical language models lack any resemblance to even a rudimentary linguistic guistic theories into a statistical framework? Why do current practical language models lack any resemblance to even a rudimentary linguistic theory? Why did 20 years of research fail to yield practical and significant imp els lack any resemblance to even a rudimentary linguistic theory? Why did 20 years of research fail to yield practical and significant improvements over the trigram, which was proposed in its essential form by Jelinek & Me I propose a few answers to these questions.

(*a*) *Linguistic theories and statistical models have different goals*

Linguistic theories deal with *existence*. They are successful if they explain (and Linguistic theories deal with *existence*. They are successful if they explain (and predict) which constructs are found in the language and which similar constructs are not. A theory is considered deficient if there are c Linguistic theories deal with *existence*. They are successful if they explain (and predict) which constructs are found in the language and which similar constructs are not. A theory is considered deficient if there are co predict) which constructs are found in the language and which similar constructs
are not. A theory is considered deficient if there are counter-examples to it. In con-
trast, statistical language models deal with *prevalen* are not. A theory is considered deficient if there are counter-examples to it. In con-
trast, statistical language models deal with *prevalence*. They are successful if they
approximate reasonably well (in log space) the p trast, statistical language models deal with *prevalence*. They are successful if they approximate reasonably well (in log space) the prevalence of the most common constructs found in the language. A model is considered de approximate reasonably well (in log space) the prevalence of the most common constructs found in the language. A model is considered deficient if there is a systematic bias, or discrepancy, between it and the phenomenon it structs found in the language. A model is considered deficient if there is a systematic
bias, or discrepancy, between it and the phenomenon it purports to describe. Thus,
a linguistic concept may be a useful tool in the co bias, or discrepancy, between it and the phenomenon it purports to describe. Thus,
a linguistic concept may be a useful tool in the context of a theory, yet prove far
less useful when it comes to improving a statistical m less useful when it comes to improving a statistical model. We have already seen an example of this in POS-based classes $(\S 2 b)$.

(*b*) *Lack of general framework*

 (b) Lack of general framework
Until recently, we have lacked a general statistical framework for incorporating
bitrary aspects of language into our models. Without such a framework accom-Until recently, we have lacked a general statistical framework for incorporating
arbitrary aspects of language into our models. Without such a framework, accom-
modating each linguistic theory involves solving a (sometime Until recently, we have lacked a general statistical framework for incorporating arbitrary aspects of language into our models. Without such a framework, accommodating each linguistic theory involves solving a (sometimes arbitrary aspects of language into our models. Without such a frame modating each linguistic theory involves solving a (sometimes hard) s
mation problem. The model described in $\S 3$ addresses this problem. mation problem. The model described in § 3 addresses this problem.
Phil. Trans. R. Soc. Lond. A (2000)

ms, and distance (non-
significance test. Over
vel of $x^2 > 15$. When

 2 si

**ATHEMATICAL,
HYSICAL
ENGINEERING
CIENCES**

OYAL

(*c*) *Mental strait-jacket of the conditional formulation*

**MATHEMATICAL,
PHYSICAL
& ENGINEERING
SCIENCES** Until recently, virtually all language modelling was done in the conditional frame-
work, i.e. by estimating $P(w | h)$. As was argued earlier, this is not conducive to
thinking about and modelling linguistic properties of t Until recently, virtually all language modelling was done in the conditional frame-
work, i.e. by estimating $P(w | h)$. As was argued earlier, this is not conducive to
thinking about and modelling linguistic properties of t Until recently, virtually all language modelling was done in the conditional framework, i.e. by estimating $P(w \mid h)$. As was argued earlier, this is not thinking about and modelling linguistic properties of the sentence as parsability). The model described in § 3 also addresses this problem. (*d*) *Impoverished priors*

Viewed within a Bayesian framework, the problem may lie in our choice of priors.
A prior is supposed to capture everything that is known about the domain before any Viewed within a Bayesian framework, the problem may lie in our choice of priors.
A prior is supposed to capture everything that is known about the domain before any data are observed. In our case, the prior should capture Viewed within a Bayesian framework, the problem may lie in our choice of priors.
A prior is supposed to capture everything that is known about the domain before any
data are observed. In our case, the prior should capture A prior is supposed to capture everything that is known about the domain before any
data are observed. In our case, the prior should capture everything that we believe
to be true about human languages in general, and about data are observed. In our case, the prior should capture everything that we believe
to be true about human languages in general, and about a specific language such as
English in particular. The very large parameter space o to be true about human languages in general, and about a specific language such as English in particular. The very large parameter space of language means that any feasible amount of training data is insufficient for overw English in particular. The very large parameter space of language means that any
feasible amount of training data is insufficient for overwhelming the prior. The choice
of prior is therefore crucial. Yet the priors we curr feasible amount of training data is insufficient for overwhelming the prior of prior is therefore crucial. Yet the priors we currently use are impoved on note take advantage of hardly anything we know about language.
As an of prior is therefore crucial. Yet the priors we currently use are impoverished: they do note take advantage of hardly anything we know about language.
As an example, consider the vocabulary clustering problem discussed i

words stand to benefit the most from clustering, yet they do not occur often enough As an example, consider the vocabulary clustering problem discussed in $\S 2 b$: rare
words stand to benefit the most from clustering, yet they do not occur often enough
in corpora for reliable automatic clustering. However words stand to benefit the most from clustering, yet they do not occur often enough
in corpora for reliable automatic clustering. However, much useful information can
be provided manually about many semantic classes, such in corpora for reliable automatic clustering. However, much useful information can be provided manually about many semantic classes, such as named entities. If such information can be encoded in a 'soft' prior, automatic c be provided manually about many semantic classes, such as named entities. If such information can be encoded in a 'soft' prior, automatic clustering methods may yet prove successful.

In summary, it could be argued that attempts to integrate linguistic knowledge into our models have so far failed because we do not yet know how to appropriately In summary, it could be argued that attempts to integrate linguistic knowledge
into our models have so far failed because we do not yet know how to appropriately
encode such knowledge, namely, how to optimally combine it w into our models have so far failed because we do not yet know how to appropriately
encode such knowledge, namely, how to optimally combine it with data. Put yet
another way, we have not figured out how to simultaneously ge encode such knowledge, namely, how to optimally combine it with data. Put yet
another way, we have not figured out how to simultaneously get the most out of
both our knowledge and our data. Between knowledge without data another way, we have not figured out how to simultaneously get the most out of both our knowledge and our data. Between knowledge without data and data without knowledge, the latter (witness the *n*-gram) is apparently mo both our knowledge and our data. Between knowled
knowledge, the latter (witness the *n*-gram) is app
is no inherent reason why we cannot have both.

Is no inherent reason why we cannot have both.
I am grateful to Ciprian Chelba, Stanley Chen, Fred Jelinek, John Lafferty, Jerry Zhu and
especially Mari Ostendorf for helpful discussions and suggestions. I am also grateful es are allowed to Ciprian Chelba, Stanley Chen, Fred Jelinek, John Lafferty, Jerry Zhu and
especially Mari Ostendorf for helpful discussions and suggestions. I am also grateful to Karen
Spärck Jones and Gerald Gazdar for v especially Mari Ostendorf for helpful discussions and suggestions. I am also grateful to Karen
Spärck Jones and Gerald Gazdar for very useful feedback on a draft of this paper.

References

Baker, J. K. 1979 Trainable grammars for speech recognition. *Speech communication papers, 97th Mtg of the Acoustic Society of America* (ed. D. H. Klatt & J. J. Wolf), pp. 547–550. Baker, J. K. 1979 Trainable grammars for speech recognition. Speech communication papers,
97th Mtg of the Acoustic Society of America (ed. D. H. Klatt & J. J. Wolf), pp. 547–550.
Beeferman, D., Berger, A. & Lafferty, J. 19

97th Mtg of the Acoustic Society of America (ed.
eferman, D., Berger, A. & Lafferty, J. 1997 A merica.
Proc. ACL-EACL '97 Joint Conf., pp. 373–380.
llogarda. J. B. 1998 A multispan language model Beeterman, D., Berger, A. & Lafterty, J. 1997 A model of lexical attraction and repulsion. In
 Proc. ACL-EACL '97 Joint Conf., pp. 373–380.

Bellegarda, J. R. 1998 A multi-span language modeling framework for large vocab

- *Proc. ACL-EACL '97 Joint Conf.*, pp. 373–3
llegarda, J. R. 1998 A multi-span language n
recognition. *IEEE Trans. SAP* 6, 456–467.
llogarda, J. R. 2000 J.argo vocabulary spooch Bellegarda, J. R. 1998 A multi-span language modeling framework for large vocabulary speech recognition. *IEEE Trans. SAP* 6, 456–467.
Bellegarda, J. R. 2000 Large vocabulary speech recognition with multi-span statistical
- recognition. *IEEE Trans. SAP* 6, 456-

llegarda, J. R. 2000 Large vocabulary s

models. *IEEE Trans. SAP* 6, 76–84.

rger A. Dolla Pietra. S. & Dolla Piet Bellegarda, J. R. 2000 Large vocabulary speech recognition with multi-span statistical language
models. IEEE Trans. SAP 6, 76–84.
Berger, A., Della Pietra, S. & Della Pietra, V. 1996 A maximum entropy approach to natural

models. *IEEE Trans. SAP* 6, 76–84.
rger, A., Della Pietra, S. & Della Pietra, V. 19
language processing. *Comp. Ling.* 22, 39–71.
own. P. E. Della Pietra, V. J. doSouza, P. V. Berger,A., Della Pietra, S. & Della Pietra, V. 1996 A maximum entropy approach to natural
language processing. *Comp. Ling.* 22, 39–71.
Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C. & Mercer, R. L. 1991 Cl

language processing. *Comp. Ling.* 22, 39–71.
own, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C. & Mer
N-gram models of natural language. *Comp. Ling.* 18, 467–478.
selba. C. & Jelinek. F. 1000 Becognition perform Brown,P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C. & Mercer, R. L. 1991 Class-based
 N-gram models of natural language. *Comp. Ling.* **18**, 467–478.

Chelba, C. & Jelinek, F. 1999 Recognition performance of a

N-gram models of natural lelba, C. & Jelinek, F. 1999
Eurospeech 4, 1567–1570. *Phil. Trans. R. Soc. Lond.* A (2000)

THE

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL
TRANSACTIONS

ATHEMATICAL,

UNXO

- Incorporating linguistic structure 1323
Chen, S. F. & Rosenfeld, R. 1999 Efficient sampling and feature selection in whole sentence maximum entropy language models. In *Proc. ICASSP, Phoenix, AZ, March 1999*, pp. 549{ 552. maximum entropy language models. In Proc. ICASSP, Phoenix, AZ, March 1999, pp. 549–
552.
Darroch, J. & Ratcliff, D. 1972 Generalized iterative scaling for log-linear models. Ann. Math.
5tat. 43, 1470–1480
- 552.
stat. 43, 1470–1480.
Stat. 43, 1470–1480.
^{Illa} Pietre. S. Delle Pi
	- Darroch, J. & Katchit, D. 1972 Generalized iterative scaling for log-linear models. Ann. Math.
 Stat. 43, 1470–1480.

	Della Pietra, S. Della Pietra, V., Mercer, R. & Roukos, S. 1992 Adaptive language modeling

	using mini Stat. 43, 1470–1480.
Illa Pietra, S. Della Pietra, V., Mercer, R. & Roukos, S. 1992 Adaptive languag
using minimum discriminant estimation. In *Proc. ICASSP*, vol. I, pp. 633–636.
Illa Pietra, S. Della Pietra, V. ⁸t. Laf
- Della Pietra, S., Della Pietra, V., Mercer, R. & Roukos, S. 1992 Adaptive language modeling
using minimum discriminant estimation. In *Proc. ICASSP*, vol. I, pp. 633–636.
Della Pietra, S., Della Pietra, V. & Lafferty, J. 1 using minimum discriminant
Illa Pietra, S., Della Pietra, V
Trans. PAMI 19, 380–393.
Adfrou, J. J. Hilliman, F. C. *Trans.PAMI* 19, 380–393.

Godfrey, J. J., Hilliman, E. C. & McDaniel, J. 1992 Switchboard: telephone speech corpus for
	- *Trans. PAMI* 19, 380–393.

	dfrey, J. J., Hilliman, E. C. & McDaniel, J. 1992 Switchboard: telephone speech corpus

	research and development. In *Proc. ICASSP, San Francisco, March 1992*, pp. 517–520.

	of D. 1997 The 1996 Godfrey, J. J., Hilliman, E. C. & McDaniel, J. 1992 Switchboard: telephone speech corpus for
research and development. In *Proc. ICASSP, San Francisco, March 1992*, pp. 517–520.
Graff, D. 1997 The 1996 Broadcast News speec
	- *Graff*, D. 1997 The 1996 Broadcast News speech and language model corpus. In *Proc. DARPA Workshop on Spoken Language Technology*, pp. 11–14. Graff, D. 1997 The 1996 Broadcast News speech and language model corpus. In *Proc. DARPA*
Workshop on Spoken Language Technology, pp. 11–14.
Iyer, R. & Ostendorf, M. 1999 Modeling long distance dependence in language: topi
	- *Workshop on Spoken Language Technology*, pp. 11–14.
 I. *R. &* Ostendorf, M. 1999 Modeling long distance dep

	vs dynamic cache models. *IEEE Trans. SAP* **7**, 30–39.

	<u>Illuse</u> F. T. 1957 Information theory and statistica Iyer, R. & Ostendorf, M. 1999 Modeling long distance dependence in language: topic mixtu
vs dynamic cache models. *IEEE Trans. SAP* 7, 30–39.
Jaynes, E. T. 1957 Information theory and statistical mechanics. *Phys. Rev.* 10
	-
	- Vertex dynamic cache models. *IEEE Trans. SAP 7*, 30–39.
Jaynes, E. T. 1957 Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630.
Jelinek, F. 1989 Self-organized language modeling for speech recognitio pres, E. T. 1957 Information theory and statistical mechanics. *Phys. Kev.* **106**, 620–630.
linek, F. 1989 Self-organized language modeling for speech recognition. In *Readings in speech*
recognition (ed. A. Waibel & K. Jelinek, F. 1989 Self-organized language modeling for speech recognition. In *Reaatings in speech*
recognition (ed. A. Waibel & K. F. Lee), pp. 450–506. San Matteo, CA: Morgan Kaufmann.
Jelinek, F. & Mercer, R. 1980 Inter
	- recognition (ed. A. Waibel & K. F. Lee), pp. 450–506. San Matteo, CA: Morgan Kaufmann.
linek, F. & Mercer, R. 1980 Interpolated estimation of Markov source parameters from sparse
data. In *Pattern recognition in practice* linek, F. & Mercer, R. 1980 Int
data. In *Pattern recognition*
Amsterdam: North Holland.
linek, E. Morialdo, B., Boulea data. In Pattern recognition in practice (ed. E. S. Gelsema & L. N. Kanal), pp. 381–402.
Amsterdam: North Holland.
Jelinek, F. Merialdo, B., Roukos, S. & Strauss, M. 1991 A dynamic language model for speech
	- recognition. In *Proc. DARPA Workshop on Speech and Natural Language, February 1991*, pp. 293–295.
	- Kneser, R. & Ney, H. 1993 Improved clustering techniques for class-based statistical language modeling. In *Proc. Eurospeech* 2, 973-976. Kneser, K. & Ney, H. 1993 Improved clustering techniques for class-based statistical language
modeling. In *Proc. Eurospeech* 2, 973–976.
Kuhn, R. & De Mori, R. 1990 A cache-based natural language model for speech recognit
	- *IIIIII III Proc. Eurospeech* 2, 9
IEEE Trans. PAMI 12, 570–583.
Forty J. D. Slooter, D. & Temperature J. D. Slooter, D. & Temperature Kunn,K. & De Mori, K. 1990 A cache-based natural language model for speech recognition.
 IEEE Trans. PAMI 12, 570–583.

	Lafferty, J. D., Sleator, D. & Temperley, D. 1992 Grammatical trigrams: a probabilistic model

	of l
	- *IEEE Trans. PAMI* 12, 570–583.
fferty, J. D., Sleator, D. & Temperley, D. 1992 Grammatical trigrams: a probabilistic model
of link grammar. In *Proc. AAAI Fall Symp. on Probabilistic Approaches to Natural Language,*
Octob *Cafferty, J. D., Sleator, D. & Temperley, D. 1992 Grammatical trigrams: a probabilistic model* of link grammar. In *Proc. AAAI Fall Symp. on Probabilistic Approaches to Natural Language, October 1992, Cambridge, MA.*
Lau,
- HYSICAL
ENGINEERING
CIENCES October 1992, Cambridge, MA.
Lau, R., Rosenfeld, R. & Roukos, S. 1993 Trigger-based language models: a maximum entropy
approach. In *Proc. ICASSP* II, 45–48. Lau, K., Kosenfeld, K. & Koukos, S. 1993 Irigger-based language models: a maximum entropy
approach. In *Proc. ICASSP* II, 45–48.
Marcus, M., Santorini, B. & Marcinkiewicz, M. A. 1993 Building a large annotated corpus of
En
	- approach. In *Proc. ICASSP* II, 45–48.
arcus, M., Santorini, B. & Marcinkiewicz, M. A. 1993 B
English: the Penn Treebank. *Comp. Ling.* 19, 313–330.
¹¹ ^{M. 1005} Markov padom fields on the graphs of pa Marcus,M., Santorini, B. & Marcinkiewicz, M. A. 1993 Building a large annotated corpus of
English: the Penn Treebank. *Comp. Ling.* **19**, 313–330.
Miller, M. 1995 Markov random fields on the graphs of natural language. In
	- *Miller, M. 1995 Markov random fields on the graphs of natural language. In Proc. Language Modeling Summer Workshop, Johns Hopkins University, August 1995.*
Modeling Summer Workshop, Johns Hopkins University, August 1995.
	- *Modeling Summer Workshop, Johns Hopkins University, August 1995.*
Price, P. 1990 Evaluation of spoken language systems: the ATIS domain. In *Proc. 3rd DARPA*
Speech and Natural Language Workshop, June 1990, pp. 91–95. Price, P. 1990 Evaluation of spoken language systems: the ATIS domain. In Proc. 3rd DARPA
Speech and Natural Language Workshop, June 1990, pp. 91–95.
R[osenfeld, R. 1996 A maximum entropy a](http://gessler.ingentaselect.com/nw=1/rpsv/cgi-bin/linker?ext=a&reqidx=/0885-2308^28^2910L.187[aid=539862,csa=0885-2308^26vol=10^26iss=3^26firstpage=187,doi=10.1006/csla.1996.0011])pproach to adaptive statistical la
	- *Speech and Natural Language Workshop, June 1990*, pp. 91–95.

	Senfeld, R. 1996 A maximum entropy approach to adaptive statistical language modeling.
 Computer Speech Language 10, 187–228. (Longer version available: Carn Computer Speech Language 10, 187–228. (Longer version available: Carnegie Mellon technical report CMU-CS-94-138.)
	- Rosenfeld, R. 1997 A whole sentence maximum entropy language model. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, CA, December 1997*, pp. 230–237. IEEE. shop on Automatic Speech Recognition and Understanding, Santa Barbara, CA, December 1997, pp. 230–237. IEEE.
Rosenfeld, R., Wasserman, L., Cai, C. & Zhu, X. 1999 Interactive feature induction and logis-
tic regression for
	- tic regression for whole sentence exponential language models. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, December 1999.*
Automatic Speech Recognition and Understanding, Keystone, CO, December 1999.
Seymore, K. & Rosenfeld, R. 1997 Using story topics for language m
	- Seymore, K. & Rosenfeld, R. 1997 Using story topics for language model adaptation. In *Proc. Eurospeech* 4, 1987–1990.

Downloaded from rsta.royalsocietypublishing.or
1324 *R. Rosenfeld* Downloaded from rsta.royalsocietypublishing.org

K. Kosenjeta
Sleator, D. & Temperley, D. 1991 Parsing English with a link grammar. Carnegie Mellon Uni-
versity Computer Science Department technical report CMU-CS-91-196. **MATHEMATICAL,
PHYSICAL
& ENGINEERING
SCIENCES** versity Computer Science Department technical report CMU-CS-91-196.
Newsity Computer Science Department technical report CMU-CS-91-196.
News A. Chelba C. Engle D. Jimenez, V. Mangu, L. Printz, H. Bistad versity Computer Science Department technical report CMU-CS-91-196.
Stolcke, A., Chelba, C., Engle, D., Jimenez, V., Mangu, L., Printz, H., Ristad, E., Rosenfeld, R.

versity Computer Science Department technical report CMU-CS-91-196.
blcke, A., Chelba, C., Engle, D., Jimenez, V., Mangu, L., Printz, H., Ristad, E., Rosenfeld, R.
& Wu, D. 1997 Structure and performance of a dependency la 5, 2775–2778.
8, Wu, D. 1997 ?
5, 2775–2778.
ard W. 1990 Tl Ward, W. 1990 The CMU air travel information service: understanding spontaneous speech. In

Proc. DARPA Speech and Natural Language Workshop, June 1990, pp. 127-129. Ward, W. 1990 The UMU air travel information service: understanding spontaneous speech. In
 Proc. DARPA Speech and Natural Language Workshop, June 1990, pp. 127–129.

Zhu, X. J., Chen, S. F. & Rosenfeld, R. 1999 Linguist

Proc. DARPA Speech and Natural Language Workshop, June 1
u, X. J., Chen, S. F. & Rosenfeld, R. 1999 Linguistic features
entropy language models. In *Proc. Eurospeech* 4, 1807–1810.

Discussion

Discussion
J. Cussens (*University of York, UK*). How does your framework compare to that
of Della Pietra *et al.*'s 'Inducing features for random fields'? *Discussion*
J. CUSSENS (*University of York, UK*). How does your frameword Della Pietra *et al.*'s 'Inducing features for random fields'?

J. CUSSENS (*University of York*, *U*A). How does your framework compare to that
of Della Pietra *et al.*'s 'Inducing features for random fields'?
R. ROSENFELD. There are two differences, relating to training and to featur tion. ROSENFELD. There are two differences, relating to training and to feature selec-
in.
In the quoted reference, the domain is modelling of the spelling of words. This
main is of moderate size and therefore Gibbs sampling can

tion.
In the quoted reference, the domain is modelling of the spelling of words. This
domain is of moderate size and therefore Gibbs sampling can be used efficiently.
However, in my own work on modelling sentences, the dom In the quoted reference, the domain is modelling of the spelling of words. This domain is of moderate size and therefore Gibbs sampling can be used efficiently. However, in my own work on modelling sentences, the domain is domain is of moderate size and therefore
However, in my own work on modelling
sampling is therefore more challenging.
Regarding feature selection, the quote $\frac{1}{\circ}$ However, in my own work on modelling sentences, the domain is far larger, and sampling is therefore more challenging.
Regarding feature selection, the quoted reference uses Kullback–Liebler distance

sampling is therefore more challenging.
Regarding feature selection, the quoted reference uses Kullback–Liebler distance
to select features from a fully specified family of features. I make use of this as well,
but place t Regarding feature selection, the quoted reference uses Kullback–Liebler distance
to select features from a fully specified family of features. I make use of this as well,
but place the emphasis on eliciting new families of to select features from a fully specified family of features. I make use of this as well,
but place the emphasis on eliciting new families of features from specialists looking
at the corpora rather than requiring a family but place the emphasis on eliciting new families of features from specialists looking
at the corpora rather than requiring a family of features to be available. The idea is
to mesh the automatic procedure with human interv

at the corpora rather than requiring a family of leatures to be available. The idea is
to mesh the automatic procedure with human intervention at the right point.
D. B. JAMES. Nouns and verbs are basic to language: why is b. B. JAMES. Nouns and verbs are basic to
noun-verb distinction is not being made? noun–verb distinction is not being made?
R. ROSENFELD. There are other, more mundane, deficiencies in the model that are

**IATHEMATICAL,
HYSICAL**
ENGINEERING
CIENCES also not dealt with. This is beyond what we can currently achieve with statistical means.

also not dealt with. This is beyond what we can currently achieve with statistical
means.
P. A. TAYLOR (*University of Edinburgh, UK*). Are the problems with trigram models
due mainly to data sparsity or to an inherent mod P. A. TAYLOR (*University of Edinburgh*, *UK*). Are the problems wit
due mainly to data sparsity or to an inherent model limitation?

P. A. TAYLOR (*University of Eainburgh*, UA). Are the problems with trigram models
due mainly to data sparsity or to an inherent model limitation?
R. ROSENFELD. Model limitations are the main problem. Using a smaller vocab R. ROSENFELD. Model limitations are the main problem. Using a smaller vocabu-
lary is infeasible since the task is to model unconstrained language. Back-off rates
are already very low, so having more data would only affect lary is infeasible since the task is to model unconstrained language. Back-off rates are already very low, so having more data would only affect a small percentage of sentences.

Sentences.

Sentences.

Sentences.

Sentences.

Sentences.

Sentences.