NAMED ENTITY TAGGED LANGUAGE MODELS

Yoshihiko Gotoh *Steve Renals* *Gethin Williams*

University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello St., Sheffield S1 4DP, UK
e-mail: {y.gotoh, s.renals, g.williams}@dcs.shef.ac.uk

ABSTRACT

We introduce Named Entity (NE) Language Modelling, a stochastic finite state machine approach to identifying both words and NE categories from a stream of spoken data. We provide an overview of our approach to NE tagged language model (LM) generation together with results of the application of such a LM to the task of out-of-vocabulary (OOV) word reduction in large vocabulary speech recognition. Using the Wall Street Journal and Broadcast News corpora, it is shown that the tagged LM was able to reduce the overall word error rate by 14%, detecting up to 70% of previously OOV words. We also describe an example of the direct tagging of spoken data with NE categories.

1. INTRODUCTION

The accurate identification of proper names and other *named entities* in spoken language is likely to be an essential component of systems performing tasks such as speech understanding, information retrieval and information extraction. Furthermore approaches based on named entity (NE) identification have the potential to improve the performance of large vocabulary speech recognition systems both through a structuring of the recogniser output (*e.g.*, as a cue to punctuation and capitalisation) and a reduction in out-of-vocabulary (OOV) words.

Current rule-based systems (*e.g.*, [1, 2]) for NE identification make use of punctuation, capitalisation and other features found in text, and would thus be difficult to apply to raw speech recogniser output — although no real effort has been made to develop such systems for spoken language. Recently hidden Markov model (HMM) based systems have been developed for NE identification with a precision/recall performance similar to that of the best rule-based systems [3]. Further, the HMM-based systems have demonstrated only a small amount of degradation when applied to speech recogniser output (after retraining) [4]. This indicates that the combination of simple finite state models and powerful estimation algorithms that has served well in speech recognition is transferable to more complex speech and language problems involving some degree of understanding.

We have developed a stochastic finite state machine structure (similar to that of [3]) for use with an acoustic model that is able to identify both words and NEs from a stream of spoken data. The approach reflects the observation (well argued in [4]) that a large number of names in the same category are stereotypical or follow simple rules (*e.g.*, personal names, names of organisations). Hence locally constrained name models for these categories can be statistically constructed given a suitably labelled corpus. Also, various lexical cues can be used for the identification of boundaries or name categories (*e.g.*, titles are often followed by personal names, location names often appear after prepositions), although we have carried out simple experiments that indicate such cues are not always reliable.

In this paper we present an overview of our approach to building NE tagged language models (LMs), together with experimental results on OOV reduction for large vocabulary continuous speech recognition (LVCSR) system using the Wall Street Journal (WSJ) and Broadcast News (BN) corpora. We also demonstrate the application of the approach to NE extraction from spoken data.

2. TAGGED LANGUAGE MODELLING

Frequently occurring NEs may be identified if a vocabulary list contains both a word and its name category information. A problem here is that the vocabulary size is limited (typically 20 to 60 thousand words) although there exist a greater number of NEs. Unfortunately many of them do not occur frequently enough to be included in a vocabulary list.

To this end a two-level finite state machine is utilised in this paper. At the first level, a locally conditioned probability model steers the generation of a sequence of name categories and words (name category information may also attributed to some words). Name categories are further extended to names in the second level. A name model is constructed for each category; it may simply be a bag of names (with each entry possibly containing multiple words), or further elaborated by using higher order n-gram type model. This two-level finite state machine is referred to as a tagged LM and formally described below.

2.1. Formulation

A tagged LM is an extension to conventional n-gram models. First, let $<w_1, \cdots, w_i>$ denote a sequence of words. Suppose there exist $L+1$ different tagged classes, $\mathcal{T} = \{t^{[0]}, t^{[1]}, \cdots, t^{[L]}\}$. $t^{[0]}$ is included for notational convenience to indicate those words not belonging to any name categories. It is assumed that each word w_i in the sequence is classified as one of the tagged classes, denoted by $t_i \in \mathcal{T}$. As a convention here, a unique identifier e_i for w_i is defined as

$$e_i = \begin{cases} <t, w>_i & \text{if } <t, w>_i \in \mathcal{V}, \\ t_i & \text{if } <t, w>_i \notin \mathcal{V}. \end{cases} \quad (1)$$

where $\mathcal{V} = \{<t, w>^{[1]}, \cdots, <t, w>^{[M]}\}$ is a set of vocabulary items with size M.

A tagged LM computes a score for each word w_i given a sequence of identifiers $e_1^{i-1} = <e_1, \cdots, e_{i-1}>$ by

$$\begin{aligned} f(w_i|e_1^{i-1}) &= \sum_{e_i \in (\mathcal{V} \cup \mathcal{T})} f(w_i, e_i|e_1^{i-1}) \\ &\sim \sum_{e_i \in (\mathcal{V} \cup \mathcal{T})} f(w_i|e_i) f(e_i|e_1^{i-1}) \end{aligned} \qquad (2)$$

In Equation (2), $f(e_i|e_1^{i-1})$ is a standard type n-gram model with a vocabulary set, $\mathcal{V} \cup \mathcal{T}$ where $\cup$ implies a union, and

$$f(w_i|e_i) = \begin{cases} 1 & \text{if } e_i = <t, w>_i \in \mathcal{V}, \\ f(w_i|t_i) & \text{if } e_i = t_i \in \mathcal{T}. \end{cases} \qquad (3)$$

where $f(w_i|t_i)$ is the unigram probability of word w_i in tagged class $t_i \in \mathcal{T}$.

2.2. Decoding Named Entities

Equations (2) and (3) may be used to estimate the language model probabilities when decoding. Alternatively, (2) may be approximated by maximisation:

$$f(w_i|e_1^{i-1}) \sim \max_{w,t \in (\mathcal{V} \cup \mathcal{T})} f(w|e) f(e|e_1^{i-1}) \ . \qquad (4)$$

This allows a decoder to recover a sequence of words and their name categories:

$$(\hat{w}_i, \hat{t}_i) \sim \operatorname*{argmax}_{w,t \in (\mathcal{V} \cup \mathcal{T})} f(w|e) f(e|e_1^{i-1}) \ . \qquad (5)$$

Note that this approach essentially performs a named entity tagging operation. An acoustic model may provide word hypotheses for w_i, then a tagged class information t_i is scored together with w_i by the two-level finite state model.

3. NAMED ENTITY TAGGED LM

The NE tagged LM was estimated from the BN text corpus. The corpus was initially processed using the tagger from the *LaSIE* (Large Scale Information Extraction) system.

3.1. The *LaSIE* Named Entity Tagger

The *LaSIE* NE tagger, developed at the University of Sheffield, is a rule-based system using list-lookup, grammar-based parsing and name coreference [1, 2]. In recent named entity evaluations in the DARPA Message Understanding Conference (MUC), it has performed at a level of around 90% precision/recall score.

The tagger recognises and classifies those classes of naming expressions, specified in the MUC NE task definition including named entities ("ORGANISATION", "PERSON", "LOCATION"), temporal ("DATE", "TIME"), and number expressions ("MONEY", "PERCENTAGE") [5]. Although these classes are not the only categories of proper names, they account for the majority of proper name occurrences in business newswire text. An example text segment, processed by the *LaSIE* system, is shown below:

- "$\underbrace{\text{O. J. Simpson}}_{\text{PERSON}}$ is traveling to $\underbrace{\text{Britain}}_{\text{LOCATION}}$ $\underbrace{\text{next month}}_{\text{DATE}}$, and he will appear on a new British TV show and address $\underbrace{\text{Oxford University's}}_{\text{ORGANISATION}}$ debating club."

from the "Broadcast News" corpus — April 1996

3.2. NE Tagged LM from Broadcast News Corpus

The BN text corpus contains data from transcribed news broadcasts, designed for use in the CSR Hub 4 Evaluation tasks. It covers the period 1992–96, with over 130 million words in the LM training set. The *LaSIE* NE tagger produced a copy of the original text with NE expressions marked up in SGML. Temporal expression and number expression tags were not used in the experiments, leaving four NE tags, labelled <PER>, <ORG>, <LOC>, and <NAM> — the latter tag being used for cases of unresolvable type ambiguity ("NAME"). The <UNK> tag was also added for those not belonging to any name categories (a set of five tags in $\mathcal{T}$).

Equations (2) and (3) provide an approach to a tagged LM generation procedure. From the SGML marked BN corpus, the most frequent 20,000 items (some words were attributed with name information) were selected for the vocabulary set $\mathcal{V}$. This resulted in total of 20,005 items in $\mathcal{V} \cup \mathcal{T}$. Then the n-gram statistics, $f(e_i|e_1^{i-1})$, were counted for $\mathcal{V} \cup \mathcal{T}$. For each tagged class in $\mathcal{T}$, the unigram statistics were counted for those words not included in $\mathcal{V}$.

The following is one example of text processing used for the LM generation. It is assumed (for demonstration purposes only) that tagged words, <PER – J.> and <ORG – University's>, were not found in $\mathcal{V}$. A sequence of words (with or without tags), marked by •, were used for n-gram model generation, then OOV words (*i.e.*, "J." and "University's") were listed on the unigram count table for each category:

- "<PER – O.> <PER> <PER – Simpson> is traveling to <LOC – Britain> next month, and he will appear on a new British TV show and address <ORG – Oxford> <ORG> debating club."

<PER>	<LOC>	<ORG>
⋮	⋮	⋮
J.		University's
⋮		⋮

The second example allows multiple-words as single vocabulary entries. In this example, it is assumed that neither multiple-word entry (*i.e.*, <PER – O._J._Simpson> nor <ORG – Oxford_University's>) was in $\mathcal{V}$, resulting in appearance in the unigram count tables:

- "<PER> is traveling to <LOC – Britain> next month, and he will appear on a new British TV show and address <ORG> debating club."

<PER>	<LOC>	<ORG>
⋮	⋮	⋮
O._J._Simpson		Oxford_University's
⋮		⋮

The first approach was used for the experiments in the next section, however the second approach may be very interesting when using with appropriate procedure for building the multiple-word pronunciations.

4. EXPERIMENTS

In this paper we have applied NE tagged LMs in two areas: the extraction of NEs from spoken data and the reduction of OOV items in large vocabulary speech recognition.

4.1. Named Entity Extraction from Speech

A speech recognition / NE extraction experiment was carried out on the BN (CSR Hub 4) task, using the ABBOT recurrent network acoustic model [6]. The acoustic model consisted of four recurrent networks, each with 53 context-independent phone classes (plus silence). The outputs of these networks, which may be interpreted as posterior probabilities, were averaged in the log probability domain. Two of the networks were trained on all of the 74 hour training set (CSR5), the other two were trained on the studio speech portions (F0 and F1 conditions) of the training data. In both cases, one network was trained on a time-reversed sequence of perceptual linear prediction feature vectors (since recurrent networks are time-asymmetric). The test set was a 173 utterance subset of the 1997 Hub 4 evaluation data with a duration of approximately 30 minutes.

The finite state nature of the statistical method (instead of using the rule-based approach of most NLP systems) may be exploited by a tight coupling with the recogniser. The NOWAY decoder is an efficient, flexible single pass LVCSR decoder developed at the University of Sheffield [7]. NOWAY is a stack-based decoder; single word extensions to decoding hypotheses are generated by the acoustic model and then scored by the language model (or the finite state machine described in Section 2). The search, acoustic model, and language model thus have a clear, decoupled interface. This enables the integration of NE tagged LMs in a single pass search to allow online NE tagging and word transcription.

Shown below is an excerpt from the reference transcription of the 1997 Hub 4 evaluation data:

- BILL CLINTON AND BOB DOLE ARE BOTH COURTING VOTERS IN THE MIDWEST TODAY THE PRESIDENT WILL SPEAK IN PARMA OHIO AND THEN MOVE ON TO DETROIT

from the "CSR Hub 4" evaluation data — 1997

By using the conventional type LM (*i.e.*, without NE tags), the decoder simply typed out the output using uppercase characters. This example failed to recover "`AND`". It also substituted "`PARMA`" and "`DETROIT`" with "`PART BY`" and "`DETROIT'S`" respectively:

- `BILL CLINTON BOB DOLE ARE BOTH COURTING VOTERS IN THE MIDWEST TODAY <SIL> THE PRESIDENT WILL SPEAK IN PART BY OHIO AND THEN MOVE ON TO DETROIT'S`

An example of the decoder output, when using the NE tagged LM, is shown below. The decoder can be programmed to produce lowercase characters except capitalisation of the detected NEs:

- `Bill Clinton Bob Dole are both courting voters in the midwest today <SIL> the president will speak in part by Ohio and then move on to Detroit's`

4.2. NE Tagged LM for OOV Reduction

This section considers a problem related to the data sparsity; that of processing OOV words relative to the conventional n-gram model. In the basic n-gram approach, a finite set of words is chosen as the vocabulary, and all other words are regarded as unknown. Typical LVCSR systems use a vocabulary of up to 65 thousand words, where the vocabulary is often chosen according to unigram frequencies in the training corpus. Such vocabularies usually cover 90–99% of words in novel data from a similar domain. Obviously, those words not included in the vocabulary cannot be recognised. Simply increasing the vocabulary size is not always an acceptable option, since it requires recomputing the n-gram model, and may substantially increase the n-gram size. Additionally there may not be sufficient data to estimate much more than unigram statistics for many words that would otherwise be OOV.

	WSJ LM	BN LM
#words with pronunciations	123,848	109,157
#pronunciations	130,964	115,286

Table 1. This table shows the total number of words and pronunciations in the pronunciation dictionary for NE tagged LMs. Some words had more than one pronunciation.

	Hub 3	Hub 4
conventional LM	250 (4.1%)	117 (2.1%)
tagged LMs		
UNK extension	76 (1.3%)	31 (0.6%)
NE extension	77 (1.3%)	31 (0.6%)

Table 2. This table shows the OOV rate for the conventional and the tagged LMs. The total number of words in the reference transcription were 6059 and 5555 words for the Hub 3 and 4 evaluation set.

Alternatively, a unigram extension (*i.e.*, $f(w_i|t_i)$ in Equation (3)) of a tagged LM may be used as a device to cover a subset of the possible vocabulary. It essentially models those words which the conventional n-gram alone would classify as OOV. As a consequence, the OOV rate is expected to be lower when using the tagged LM. It also restricts the size of the model to manageable level and effectively pools the n-gram statistics of semantically related words. A large proportion of OOV words are proper names. By modelling these words, it seems likely that speech recognition performance on such a task could be improved.

Speech recognition experiments were carried out for

- NE tagged WSJ corpus / 1995 Hub 3 evaluation data.
- NE tagged BN corpus / 1997 Hub 4 evaluation data.

The WSJ corpus (also NE tagged by the *LaSIE* system) covers the period 1987–94 with over 100 million words, from which most frequent 19,952 items were selected as the vocabulary. The 1995 Hub 3 evaluation test set (C0 condition) consists of roughly 45 minutes of speech. The NE tagged BN LM (with 20,000 vocabulary items) and the 1997 Hub 4 evaluation data were described earlier.

Three sets of LMs were constructed for each NE tagged corpus; a conventional type LM and two tagged LMs with UNK extension / NE extension. The NE extension consisted of five labels — <PER>, <ORG>, <LOC>, <NAM>, and <UNK>. The UNK extension simply summarised all entries from these five categories under one label <UNK>, implying no category information available. After discounting and smoothing, the resulting models contained about 8 million trigrams. A breakdown of the pronunciation dictionaries for the NE tagged LMs are shown in Table 1.

The OOV rate of the test set was evaluated by comparing the reference transcription with the pronunciation dictionaries. There were a total of 6059 (Hub 3) and 5555 words

	Hub 3	Hub 4	
		medium	*narrow*
conventional LM	20.5	34.7	44.2
tagged LMs			
(flat estimate)	(18.2)	(33.3)	(42.9)
UNK extension	17.7	-	38.4
NE extension	17.7	-	38.5

Table 3. This table shows the WER (%) for conventional and tagged LMs. "*medium*" and "*narrow*" for the Hub 4 evaluation set refer to the search beam width used by the speech recogniser. For the flat estimate variation, $P(w|e)$ was set to 10^{-5} since there were about 10^5 items in $\mathcal{T}$. Due to time constraints, tagged LM results for a *medium* beam decoding of the Hub 4 data could not be obtained.

(Hub 4) in the transcription; 250 (4.1%) and 117 (2.1%) words, respectively, were not in the trigram vocabulary $\mathcal{V}$. For each case, about 70% of these OOVs were recovered by the tagged unigram word set, reducing the effective OOV rate to 1.3% and 0.6% respectively (Table 2).

Table 3 shows the word error rate (WER). Hub 3 evaluation data and Hub 4 with *medium* and *narrow* search beam widths were tested. A variation to the tagged LM with UNK extension was also tested using a flat estimate of $P(w|e)$. All tagged LMs (UNK and NE extensions, flat estimate) have shown useful level of improvement in recognition performance. It is not surprising that tagged LMs improved the performance over the conventional model. For both Hub 3 evaluation data and Hub 4 with *narrow* beam, 70% reduction of OOV rates by tagged LMs were directly translated into about a 14% reduction in WER. Although there was no significant difference in WER between tagged LMs with UNK extension and NE extension, there was an improvement in performance by using an estimated unigram model for $P(w|e)$ over a flat estimate.

Although there is little evidence that a tagged LM derived from semantically marked corpus is better than blind approach (of mapping all to the UNK symbol), the former is preferred over the blind approach as the former not only improves the recognition performance but identifies the names from speech data. Furthermore, more sophisticated approaches for NE tagged LMs may lead to further improvement if the search is appropriately constrained by semantic information.

5. POTENTIAL BENEFITS

In the experiments, we have used NE tagged LMs (locally constrained stochastic finite state automata) in two areas:

1. extraction of named entities from spoken data.
2. reduction of OOV items in LVCSR system.

NE extraction (*item 1*) will have a strong impact in many application areas such as spoken data retrieval system (because extracted NEs are likely to be important key words). LVCSR systems may benefit from the approach because NEs can be used as cues to punctuation and capitalisation. Experiments involving the evaluation of NE identification performance are underway, as part of our participation in the NE spoke of 1998 CSR Hub 4 evaluation. Additionally, tagged LM approach is a computationally efficient way to recover from recognition errors caused by OOVs and the LVCSR experimental results have shown useful level of improvement (*item 2*).

There exist further potential benefits from the approach:

3. accommodating up-to-date topics in an LM.
4. prediction of proper names / OOV items.
5. finite state machine based NE tagger.

It is often observed that, even for the same task domain (*e.g.*, broadcast news material), topic changes very rapidly (thus LM data might be soon obsolete, *e.g.*, Viagra, Lewinsky). We may be able to accommodate up-to-date topics assuming that extracted NEs indicate the topic for newer materials (*item 3*). Proper name prediction (*item 4*) is in conjunction with our on-going work on confidence measures.

We are also implementing a finite state machine based NE tagger (*item 5*). In comparison to rule-based approaches, the locally constrained finite state machine is unable to cope with global information that might be uncovered by use of a parser and coreferencing. However, it also has many advantages. Firstly, a state machine structure meshes with a conventional LVCSR system and secondly, considerable speed-up is expected because it does not make use of a parser or coreference. Thirdly, and probably most importantly, it does not assume any specific task domain or language because it is statistical and corpus based.

ACKNOWLEDGEMENT

This work is funded by "SPRACH" and "THISL" (ESPRIT 20077/23495). Gethin Williams is supported by an EPSRC studentship. The authors wish to express their great appreciation to Rob Gaizauskas, Hamish Cunningham, and Kevin Humphreys (University of Sheffield, Computer Science) for their support in using the *LaSIE* NE tagger.

REFERENCES

[1] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 207–220, Maryland, November 1995.

[2] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[3] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*, pages 194–201, Washington, DC, April 1997.

[4] Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. Named entity extraction from speech. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[5] Defense Advanced Research Projects Agency. Proceedings of the sixth message understanding conference (MUC-6). Maryland, November 1995.

[6] Anthony J. Robinson. The application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.

[7] Steve Renals and Mike Hochberg. Efficient evaluation of the LVCSR search space using the NOWAY decoder. In *Proceedings of ICASSP-96*, volume 1, pages 149–152, Atlanta, May 1996.