INTEGRATED TRANSCRIPTION AND IDENTIFICATION OF NAMED ENTITIES IN BROADCAST SPEECH

Steve Renals *Yoshihiko Gotoh*

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

ABSTRACT

This paper presents an approach to integrating functions for both transcription and named entity (NE) identification into a large vocabulary continuous speech recognition system. It builds on NE tagged language modelling approach, which was recently applied for development of the statistical NE annotation system. We also present results for proper name identification experiment using the *Hub-4* evaluation data.

1. INTRODUCTION

The accurate identification of proper names and other *named entities* (NEs) has a useful role to play in spoken language processing, as component in speech understanding systems, and as a way of structuring recogniser output (*e.g.*, as a cue to punctuation and capitalisation).

Recently trainable hidden Markov model systems for NE identification have been reported with a precision/recall performance similar to that of the best grammar based systems and only a small amount of degradation when applied to speech recogniser output [1, 2]. We have previously presented an NE tagged language modelling approach that uses named entities to extend the recogniser vocabulary in a straightforward way [3].

In this paper we outline a statistical NE annotation system and present results for proper name identification experiments using the *Hub-4* evaluation framework. We then describe how this approach enables an integration of transcription and NE identification in a large vocabulary continuous speech recognition (LVCSR) system.

2. NAMED ENTITY ANNOTATION SYSTEM

An n-gram based NE annotation system for speech transcription was developed initially for the identification of named entities in news broadcasts, the *IE-NE* spoke of the 1998 DARPA/NIST *Hub-4* evaluation. This system consisted of an NE tagged language model (LM) and a statistical NE tagger working in pipeline with the output of a conventional LVCSR system [4]. A complete description of the NE tagged LM is provided in [3]; technical details for the development and the annotation procedure are presented in [5]. The official results for our participation in the *Hub-4* evaluation may be found in [6].

2.1. NE Tagged LM

The basic idea of the NE tagged LM is to use NE tags as categories in a class-based n-gram language model [3]. This enables the construction of extensible vocabulary speech recognition systems, along with the identification of named entities in spoken language. It is derived from a corpus marked with named entities. The vocabulary is split into two sets, the *core vocabulary* (typically the 20 000 to 65 000 most frequent words), and an *extension vocabulary* (which may be much larger) made up of words out of core vocabulary.

An NE tagged LM is an extension to conventional n-gram models; a backed off model using the core vocabulary is built over the set of words attributed with their name category information. Further, extension vocabulary words are identified with their categories and these categories are also members of the backed off model vocabulary. A separate unigram process is used to map from each name category to extension vocabulary elements.

Formally, let $< w_1, \cdots, w_i >$ denote a sequence of words. Suppose there exist $L+1$ different tagged classes, $\mathcal{T} = \{t^{[0]}, t^{[1]}, \cdots, t^{[L]}\}$. For simplicity in technical terminology, tags $\{t^{[1]}, \cdots, t^{[L]}\}$ are referred to as *name categories*. $t^{[0]}$ implies a <not_a_name> class, indicating 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 tag-word token e_i for w_i is defined as

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

where $\mathcal{V} = \{<t, w>^{[1]}, \cdots, <t, w>^{[M]}\}$ is a set of core vocabulary items with size M. This implies that the same two words having different tags are considered to be different items in the vocabulary.

Two stochastic processes are then defined: an n-gram model over tag-word tokens and a unigram extension relating words to tags. We compute a score for each word

w_i given a sequence of tokens $e_1^{i-1} = <e_1, \cdots, e_{i-1}>$ by

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

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

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

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

This model may be interpreted as a discrete HMM, in which the set of states is defined by the observed set of tag-word pairs $<t, w>$ in $\mathcal{V}$, appended with the set of tags t in $\mathcal{T}$ (to cover the extension vocabulary). The observations of the model are words. Tag-word states have a delta distribution with the relevant word being emitted with probability 1. The output distribution for tag states is a unigram distribution over the extension vocabulary, estimated from training data.

2.2. NE Identification from Speech Transcription

The most probable sequence of named entities may be identified by tracing the Viterbi path across the tag-word trellis. This search is based on n-gram relations (and possibly backed off to lower order n-grams). Further, the following constraints were used:

1. Transitions to/from out-of-vocabulary (OOV) items in name categories are prohibited[1]. This does not apply to OOVs in <not_a_name> class.

2. Transitions to/from in-vocabulary items in name categories are favoured[2]. This does not apply to in-vocabulary items in <not_a_name> class.

The first constraint improves the *precision* of NE identification. It eliminates any chance that a word might be correctly marked even if that tag-word pair does not exist in the language model. However, without this exception rule, the number of incorrect markings increases because of unbalanced sizes of name classes[3]. The second constraint was introduced after the 1998 *Hub-4* evaluation. It improved the *recall* rate by encouraging in-vocabulary names identified in each name category.

[1]For example, consider a word "GEORGE" in speech transcription. Without the first constraint, some OOV item, *i.e.*, "unknown word" in, say, <date> category might be chosen (which, of course, is not correct) because it often has higher probability than "GEORGE" in <person> name category.

[2]The same example, but this goes to the opposite direction. Without the second constraint, "unknown word" in <not_a_name> category might be chosen instead of "GEORGE" in <person> name category.

[3]In news broadcasts, <organisation>, <person>, and <location> names occurred orders of magnitude more than other temporal and number expressions.

evaluation data	hand transcription			recogniser output		
	R	*P*	*P&R*	*R*	*P*	*P&R*
1997 *Hub-4*	.70	.90	.79	.58	.75	.65
1998 *Hub-4*	.78	.90	.83	.66	.79	.72

Table 1: This table shows identification scores for proper name expressions (*i.e.*, summary of <organisation>, <person>, and <location>) on the 1997 and 1998 *Hub-4* evaluation data. The *IE-NE* scoring pipeline package "ieeval" (developed by SAIC and distributed by NIST) was used for scoring the NE annotated hypothesis, then results for proper name expressions were extracted. *R*, *P*, and *P&R* denote recall, precision, and combined precision/recall scores respectively.

2.3. Results for Proper Name Identification

We tested the statistical NE tagger on proper name identification experiments using the North American Broadcast News (BN) task. An NE tagged trigram LM was estimated from the 1996 BN text corpus (both training and test data — 150 million words)[4]. NE annotation on the corpus was done automatically using the *LaSIE-II* system [7]. *LaSIE-II* is a grammar based information extraction system developed at the University of Sheffield, which achieved over 90% combined precision/recall score on the *MUC-7* business newswire NE identification task. However the text style for news broadcasts is somewhat different, hence resulting in some errors being included in the language model.

When generating an NE tagged LM, proper name classes (<organisation>, <person>, <location>) and number expressions (<money>, <percentage>) were modelled; temporal expressions (<date>, <time>) were not considered due to a change of specification between the *MUC-7* and *Hub-4* evaluations. The derived model used a 65 000 word core vocabulary, resulting in 4.3 million bigrams and 12.9 million trigrams, together with an 85 000 word extension vocabulary.

Table 1 shows identification scores for proper name expressions on the *Hub-4* evaluation data (hand and recogniser transcriptions). Hand verified transcriptions may be considered as ones with 0% word error rate (WER). For recogniser transcriptions, we used outputs from the 1997 CU-CON system (27% WER) [8] and the 1998 SPRACH recogniser (21% WER) [4].

For hand transcriptions, the precision score reached 90% with a recall of 70%, resulting in a combined precision/recall score[5] of about 80%. For recogniser outputs, the scores declined by around 17% (relative) for

[4]For the 1998 DARPA/NIST *Hub-4* evaluation *IE-NE* spoke, we also used transcripts of *Hub-4* acoustic training data (one million words) and the 1998 North American News corpus (133 million words). An NE tagged LM was estimated for each of three data sets, speech transcriptions were marked with named entities according to each LM, then merged to produce a single and final hypothesis. More details are provided in [5, 6].

[5]A combined precision/recall score is also known as the *F-measure*

the 1997 data (27% WER) and 13% for 1998 data (21% WER). A linear relationship between the WER and the NE identification scores was observed in the 1998 *Hub-4* evaluation [9].

NE annotation errors[6] are analysed as follows:

- Most correctly identified named entities were identified using bigram or trigram constraints around each named entity (*i.e.*, a named entity itself and words before/after that named entity). When the language model was forced to back-off to unigram statistics, a bigram of an "unknown word" in <not_a_name> category followed by some other word was often more probable than the unigram of the tagged word.

- Multiple word named entities were not explicitly handled in the NE tagged LM. Post-corrections were made using which mapped a sequence of consecutive words marked with the same name tag to a single named entity. This approach was adequate for many cases (*e.g.*, "<person>BILL CLINTON"), but failed to handle cases of consecutive tags of the same type: *e.g.*, "<location>SIMI VALLEY" followed by "<location>CALIFORNIA" was incorrectly identified as "<location>SIMI VALLEY CALIFORNIA".

- Inaccuracy in automatic annotation on the training corpus caused another type of error. Occasionally the *LaSIE-II* system marked the training corpus with <name> tags when an unresolvable type ambiguity occurred between <organisation>, <person>, and <location>.

The results in Table 1 contains recent developments since the 1998 *Hub-4* benchmark test[7]. In particular, the second constraint in search process (*i.e.*, favouring transition to/from in-vocabulary items in name categories) resulted in a 2–3% improvement in recall without sacrificing precision. This constraint is a boost especially when the LM size is small because a smaller LM will result in a fewer bigram or trigram hits when decoding the tag-word trellis.

2.4. Other Statistical NE Annotation Systems

The n-gram approach presented in this paper resulted in precision and recall scores that were 5–10% worse than those reported by Miller *et al.* [9] and Palmer *et al.* [10]. Those systems were trained using only a one million word training set of manually annotated data. Ignoring technicalities, their methods modelled transitions to the current word and class, conditioned on the previous word and class: *i.e.*, transitions between classes were explicit. In contrast, we have constructed an n-gram model directly on word to word transitions, with class information treated as a word attribute. This is a serious drawback of the direct n-gram approach. As described above, the successful recovery of name expressions is heavily dependent on existence of higher order n-grams in the model. A possible way to improve the direct n-gram approach seems to be via the incorporation of constraints on a class level.

3. INTEGRATION

A problem with most current LVCSR systems arises from the unstructured nature of the recogniser output — for example the lack of punctuation and capitalisation. In this section we describe the first step to structure a speech transcription by integrating NE identification with the LVCSR system. Specifically we use identified named entities as a cue to capitalisation. We demonstrate the approach by showing the speech transcription with case information.

We have integrated the NE tagged language modelling approach into the single pass NOWAY decoder [11] used in the ABBOT/SPRACH system [4]. NOWAY is a start-synchronous stack-based decoder that operates in a single pass, using a variety of pruning techniques. A key feature of the decoder design is a clean decoupled interface between the language model and the acoustic model: the language model simply returns a probability for a word given its preceding context (which may be arbitrarily long). The same constraints described in section 2 may be applied in the decoder. Thus the implementation of more general finite state models (including class-based and NE tagged LMs) is straightforward. Integrating the NE tagged LM with the search makes possible the use of name category information to further constrain the search. Furthermore, the NE tagged LM may be used to extend the vocabulary of the recogniser without forcing recomputation of the language model.

Integrated transcription and NE identification is a key step on the way to a more structured recogniser output. Many languages, including English, use case information to identify proper names. Shown below is an excerpt from the reference transcription of the 1997 *Hub-4* evaluation data:

- "FRESH FROM HIS SUCCESS IN HELPING TO MANAGE BILL CLINTON'S RE ELECTION CAMPAIGN GEORGE STEPHANOPOULOS HAS GOT A NEW MISSION HE'S GOING TO CROSS THE ATLANTIC TO HELP BRITAIN'S OPPOSITION LABOR PARTY IN ITS BID FOR POWER"

By using the conventional type LM (*i.e.*, without NE tags), the decoder simply typed out the output without any case information:

(*e.g.*, *MUC-7*). A standard calculation:

$$P\&R = \frac{2 \cdot R \cdot P}{R + P}$$

was used in the experiment.

[6] Further details using graphs and examples from the annotated transcriptions are provided in [5].

[7] Complete results (for all name categories and for all conditions) for 1998 *Hub-4* benchmark test will be found in [6].

- "FRESHMAN SUCCESS IN HELPING TO MANAGE BILL CLINTON'S RE ELECTION CAMPAIGN GEORGE STEPHANOPOULOS HE'S GOT A NEW MISSION <SIL> IS GOING ACROSS THE ATLANTIC TO HELP BRITAIN'S OPPOSITION LABOR PARTY IN ITS BID FOR POWER"

This example contains a few transcription errors (correct transcripts in parentheses): "FRESHMAN" (FRESH FROM HIS), "HE'S" (HAS), "IS" (HE'S), and "ACROSS" (TO CROSS). Using integrated transcription and NE identification, we may both identify names and use them as a cue to capitalization:

- "freshman success in helping to manage [Bill Clinton]'s re election campaign [George Stephanopoulos] he's got a new mission <SIL> is going across the [Atlantic] to help [Britain]'s opposition [Labor Party] in its bid for power"

This approach may be used in combination with a punctuation aware language model [12].

4. SUMMARY

In this paper we have discussed a named entity tagged language model which may be integrated to a single pass recogniser to enable structured transcription. The NE tagged LM was constructed within the framework of the conventional n-gram language modeling approach. As a consequence, it was straightforward to integrate function for speech transcription and NE annotation into a single system.

We have presented results for proper name identification experiment using the *Hub-4* evaluation data. Due to the sparsity of the state space of the NE tagged LM, precision and recall levels were not as high as those approaches that directly model NE class level constraints. Further work will include the incorporation of explicit transitions between classes.

Acknowledgements. This work was funded by ESPRIT Long Term Research project SPRACH(20077) and UK EPSRC grant GR/M36717.

REFERENCES

[1] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learning name-finder," in *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*, (Washington, DC), pp. 194–201, April 1997.

[2] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, VA), February 1998.

[3] Y. Gotoh, S. Renals, and G. Williams, "Named entity tagged language models," in *Proceedings of ICASSP-99*, vol. I, (Phoenix), pp. 513–516, March 1999.

[4] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams, "An overview of the SPRACH system for the transcription of broadcast news," in *Proceedings of DARPA Broadcast News Workshop*, (Herndon, VA), February 1999.

[5] Y. Gotoh and S. Renals, "Statistical annotation of named entities in spoken audio," in *Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio*, (Cambridge), pp. 43–48, April 1999. Available from http://svr-www.eng.cam.ac.uk/~ajr/esca99/.

[6] S. Renals, Y. Gotoh, R. Gaizauskas, and M. Stevenson, "Baseline IE-NE experiments using the SPRACH/LaSIE system," in *Proceedings of DARPA Broadcast News Workshop*, (Herndon, VA), February 1999.

[7] 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 7th Message Understanding Conference (MUC-7)*, 1998.

[8] G. D. Cook and A. J. Robinson, "The 1997 ABBOT system for the transcription of broadcast news," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, VA), February 1998.

[9] D. Miller, R. Schwartz, R. Weischedel, and R. Stone, "Named entity extraction from broadcast news," in *Proceedings of DARPA Broadcast News Workshop*, (Herndon, VA), February 1999.

[10] D. D. Palmer, J. D. Burger, and M. Ostendorf, "Phrase language models for named entity tagging," in *Proceedings of DARPA Broadcast News Workshop*, (Herndon, VA), February 1999.

[11] S. Renals and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, 1999. To appear.

[12] D. Beeferman, A. Berger, and J. Lafferty, "CYBERPUNC: A lightweight punctuation annotation system speech," in *Proceedings of ICASSP-98*, vol. 2, (Seattle), pp. 689–692, May 1998.