STATISTICAL ANNOTATION OF NAMED ENTITIES IN SPOKEN AUDIO

Yoshihiko Gotoh *Steve Renals*

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

ABSTRACT

In this paper we describe stochastic finite state model for named entity (NE) identification, based on explicit word-level n-gram relations. NE categories are incorporated in the model as word attributes. We present an overview of the approach, describing how the extensible vocabulary model may be used for NE identification. We report development and evaluation results on a North American Broadcast News task. This approach resulted in average precision and recall scores of around 83% on hand transcribed data, and 73% on the SPRACH recogniser output. We also present an error analysis and a comparison of our approach with an alternative statistical approach.

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 through a structuring of the recogniser output (*e.g.,* as a cue to punctuation and capitalisation).

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 and only a small amount of degradation when applied to speech recogniser output [1, 2]. In [3], we presented a stochastic finite state machine structure for use with an acoustic model that is able to identify both words and named entities from a stream of spoken data. In this paper we describe how this framework was used for NE identification in broadcast audio, and present experiments performed within the DARPA/NIST *Hub-4E* benchmark *IE-NE* spoke.

2. FRAMEWORK

2.1. Tagged Language Modelling

The basic idea of the NE tagged LM is to use NE tags as categories in a class-based n-gram language model. This enables the construction of extensible vocabulary speech recognition systems, along with the identification of named entities in spoken language. An NE tagged LM is derived from a corpus marked with named entities. It is a backed off n-gram model with the vocabulary entries being the most frequent words attributed with their name category information. Because many proper names do not occur frequently enough to be listed in the n-gram model vocabulary, unigram extensions for less frequent names are attached in order to increase the overall vocabulary size.

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

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

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 $t_i \in \mathcal{T}$. Note that this model may be regarded as a discrete HMM, having states e_i and observations w_i.

Generally, a sufficient amount of text data is required in order to construct a statistical language model. In the experiments, we used multiple corpora marked up with named entity information (either manually or automatically using the *LaSIE-II* system). A construction procedure for the NE tagged LM is outlined in [3] using a simplified example.

2.2. Statistical Identification of Named Entities

Named entities can be identified using the NE tagged stochastic finite state model. Input to the statistical NE tagger is textual data or a transcription of spoken data, the latter typically being produced by a speech recognition system.

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 Viterbi decoder to recover a sequence of words and their name categories, *i.e.,*

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

An acoustic model provides word hypotheses for w_i, then tagged class information t_i is scored together with w_i by the finite state model.

If the sequence of words $<w_1, \cdots, w_i>$ is known, the procedure is reduced to a state machine which essentially performs a statistical NE marking operation on that word sequence.

3. EXPERIMENTS

We have performed NE identification experiments using North American Broadcast News, in the context of the 1998 NIST *Hub-4E* benchmark *IE-NE* spoke. The system consisted of three NE annotated text sources, an n-gram based NE tagger to mark speech transcripts, and other pre and post processing tools. It was built using the 1997 evaluation data as a development set, then applied to the 1998 task. The annotation categories were named entities (<organisation>, <person>, <location>), temporal (<date>, <time>), and number expressions (<money>, <percentage>).

3.1. NE tagged LMs

Three NE tagged and backed off trigram LMs were produced first, each with an independent vocabulary set plus unigram extensions:

H4-train LM was derived from transcripts of Broadcast News (BN) acoustic training data (1996-97) — approximately one million words with manual NE annotations. 18k trigram vocabulary (*i.e.,* tag-word tokens), with a further 4k vocabulary in unigram extensions.

BN96 LM was estimated from 1996 BN text corpus for training/test data — 150 million words with automatic NE annotations. 65k trigram vocabulary, with a further 85k vocabulary in unigrams.

NA98 LM was trained on a part of the 1998 North American News (NA News) corpus (1996-98 Associated Press, 1997-98 LA Times/Washington Post) — 133 million words with automatic NE annotations. 65k trigram vocabulary, with a further 145k vocabulary in unigrams.

Manual NE annotations were provided by MITRE and BBN (through NIST) and they conformed with the *Hub-4E IE-NE* task specification [4]. Automatic annotations were achieved using the *LaSIE-II* system [5]. Since *LaSIE-II* was developed according to the *MUC-7* NE task specification [6], relative time expressions were also marked for the BN and NA News corpora, conflicting with the *Hub-4E* specification.

These three sources were prepared for LM production. First, we assumed that transcripts of BN acoustic training data were the closest fit to the task domain for *Hub-4E* evaluation and that their NE annotation was reliable because it was done manually. We expected the H4-train LM from this data source would provide a baseline performance for the experiment. A problem of the H4-train transcripts is that the size (approximately one million words) is small for large vocabulary speech recognition. The 1996 BN text corpus was annotated with NE class markers to compensate this issue. It has sufficient size for building a large vocabulary system and conforms with the *Hub-4E* task domain (but probably transcription was done less carefully than for the acoustic training data). We also annotated a part of the 1998 NA News corpus. It contained a large amount of newswire texts but we expected their style to be different from the *Hub-4E* task domain[1]. Nevertheless we used this corpus because it contained more recent topics and names not available from the 1996 corpus.

3.2. System Development

To develop the NE annotation system, we used a **development data set** consisting of the 1997 *Hub-4E* evaluation data, selected from news shows broadcasted between October and November, 1996 (approximately three hours of speech, with over 30k words).

[1] Although the NA News corpus might have different text style from the *Hub-4E* task domain, the *LaSIE-II* system would achieve better annotation than for the BN corpus because it was developed for the North American business newswire as a primary target.

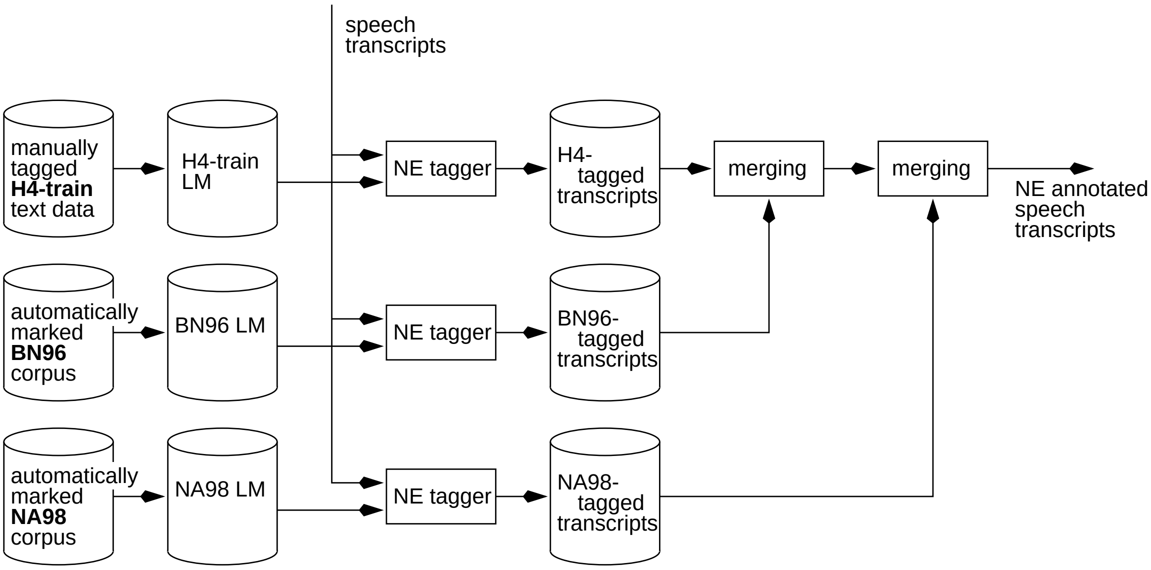

Figure 1: The statistical NE annotation system. Speech transcripts are marked with NE tags using each of three individual NE tagged LMs. Three sets of NE annotated transcripts are then merged with priority on the H4-tagged transcripts.

The statistical NE annotation system is illustrated in Figure 1. After several trial runs using the development set, an NE annotation procedure was settled as follows:

1. Mark speech transcripts with NE tags using each of three individual NE tagged LMs, resulting three sets of NE annotated speech transcripts (referred to as H4-tagged, BN96-tagged, and NA98-tagged transcripts).

2. Merge the BN96-tagged transcripts to the H4-tagged transcripts with priority on the latter. More specifically, extract NE tagged items from the BN96-tagged transcripts and, for each item, copy to the H4-tagged transcripts if any part of it is not marked. Because of the specification conflicts, temporal expression tags (*i.e.*, <date> and <time>) initially marked on the BN96-tagged transcripts were ignored at this stage.

3. Merge the NA98-tagged transcripts to the merged transcripts at Step 2, with priority on the latter. Again temporal expression tags from the NA98 LM are ignored.

4. Apply final cosmetic fixing on the merged speech transcripts.

The initial marking on speech transcripts was done using the trigram constraints described in Section 2 with one exception: when tracing the Viterbi path across the tag-word trellis, we removed the possibility of transitions to/from any out-of-vocabulary (OOV) item in each name class (*e.g.*, a word "GEORGE" was OOV with respect to, say, a <date> category) from consideration. This was regrettable because it eliminated any chance that a word might be correctly marked even if that tag-word pair did not exist in the language model. Without this exception rule, however, the number of incorrect markings increased greatly because of unbalanced sizes for tag classes (temporal and number expressions occurred an order of magnitude less than other name classes). Occasionally, a tag class of smaller size was favoured when the probability mass for the unknown in that class was very large in comparison to the probability for correct tag-word pair (*i.e.*, "GEORGE" might be marked as <date> instead of <person>).

Because this n-gram based NE tagger did not explicitly handle multiple word named entities, we made post-corrections according to a simple rule: suppose multiple and consecutive words were all marked with the same name tag, then we assumed they belonged to one named entity. For example, suppose "BILL" and "CLINTON" were both marked as <person>, then

<person>"BILL CLINTON"

became a single hypothesis. This approach, of course, had a critical side effect: "SIMI VALLEY CALIFORNIA" were marked with a single NE tag, <location> (and many such examples existed).

During the process of system development we recognised that the merging strategy (*i.e.*, priority on the H4-tagged transcripts; elimination of temporal expression tags from the BN96-tagged and the NA98-tagged transcripts) resulted in rather poor scores for temporal expressions in comparison to other name categories. Thus we prepared a post processor that mechanically marked <date> tags on the transcripts if days of a week, months (except "MARCH", "APRIL", "MAY", and "JUNE"), and four season names were not initially marked. This operation pushed up the recall score for

LM	1997 hand verified transcription			1997 CU-CON recogniser output		
	R	*P*	*P&R*	*R*	*P*	*P&R*
H4-train	.46	.84	.60	.41	.74	.53
BN96	.73	.70	.72	.62	.60	.61
NA98	.69	.67	.68	.59	.59	.59
"all"	.78	.84	.80	.66	.71	.68

Table 1: NE identification scores on the development set (1997 *Hub-4E* evaluation data). Results are shown for each of three component LMs (defined in the text) along with the merged system ("all"). *R*, *P*, and *P&R* denote recall, precision, and precision&recall scores respectively.

<date> from .58 to .90 and improved overall precision& recall score by nearly 1% (1997 hand verified transcription case).

Results for the development set. The development set contained (a) manually verified (by NIST) transcriptions and (b) transcriptions produced using the 1997 CU-CON speech recognition system [7]. The word error rate (WER) for the latter was approximately 27%. Once these transcripts were marked with NEs, they were scored against NE-annotated reference transcripts. The performance was measured using recall (*R*), precision (*P*), and a combined precision&recall score (*P&R*)[2].

For each of three individual LM sets, Table 1 shows NE identification results on the development set. This table indicates that the H4-train LM, obtained from the limited amount of manually annotated training data, resulted in a much higher precision than the other two, but had a poor recall owing to its limited vocabulary. The LMs trained on the automatically annotated data resulted in lower precision NE tagging but a higher recall score. Table 1 also shows the results for the merged system. On hand transcriptions, merged results did not reduce the precision but significantly improved the recall; on the 27% WER transcriptions, merging did result in a slightly reduced precision with respect to the H4-train model, but again gave a significant improvement in recall and a combined precision&recall score.

For hand verified transcriptions, Figure 2 shows the effect of merging the BN96-tagged and the NA98-tagged transcripts to the H4-tagged transcripts. It is observed that recall rate has improved at each stage of merging. Although text style for the 1998 NA News corpus was different, it still was able to provide measurable improvement in recall score because it contained recent name expressions.

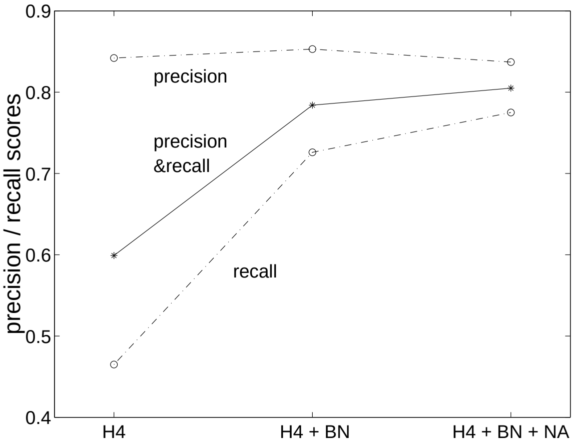

Figure 2: Effect of merging the BN96-tagged and the NA98-tagged transcripts to the H4-tagged transcripts. At each stage, precision / recall scores are shown on hand verified transcriptions from the 1997 *Hub-4E* evaluation data.

3.3. Error Analysis

In order to gain further insight for the n-gram based approach, we have broken down NE identification results to each name subclass. Figure 3 shows recall and precision scores produced using the H4-train, the BN96, and the NA98 LMs. The merged system (not shown) was a "summary" of all three and did not provide more useful information.

In general, annotated transcripts from the BN96 and the NA98 LMs achieved very similar for both recall and precision scores. It is an interesting result because their text styles were slightly different (one from news broadcast, and the other from newswire) although their LM sizes were approximately the same. In the following, we analyse NE annotation errors by closer inspection to the mark-ups on the development data set.

Recall scores. Name categories, <location> (38.6% of total NE occurrences in the annotated reference), <person> (28.3%), and <organisation> (22.3%) dominated the temporal and number expressions. Recall scores for <location> and <person> were substantially higher by the BN96-tagged and the NA98-tagged transcripts than by the H4-tagged transcripts.

The initial marking on speech transcripts was done solely using the backed off trigram relation. By inspection of annotated transcripts, it was found that most correctly marked NEs were identified through bigram or trigram constraints around each NE (*i.e.*, the NE itself and words before/after that NE). When the LM was forced to back-off to unigram statistics, the LM often estimated a bigram of an unknown word (with no

[2] A combined precision&recall score is also known as the *F-measure* (*e.g.*, *MUC-7*). A standard calculation:

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

was used in the experiment.

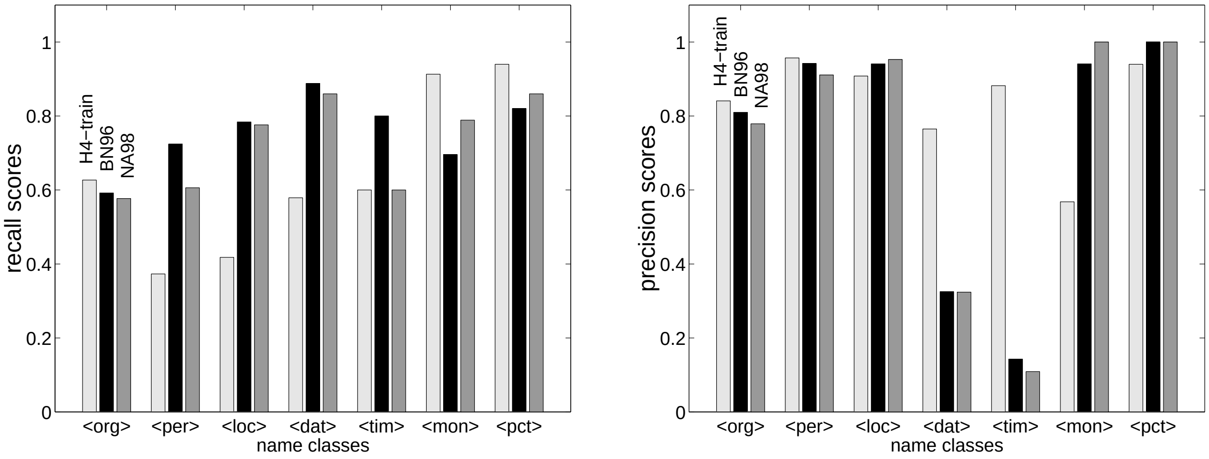

Figure 3: Comparison of NE identification results produced using the H4-train, the BN96, and the NA98 LMs, for the 1997 hand transcriptions. Recall and precision scores are shown for each name subclass. There were 1891 NE occurrences in the annotated reference distributed as: <organisation>:22.3%; <person>:28.3%; <location>:38.6%; <date>:5.7%; <time>:1.3%; <money>:1.2%; <percentage>:2.6%.

tag) followed by some other word, rather than the unigram of a tagged word[3]. Larger LMs were more likely to include the required bigrams and trigrams: thus it is not very surprising that the recall score using the H4-train LM (uni/bi/trigram: 19k, 96k, 86k entries) was less than the BN96 LM (65k, 4.3M, 12.9M entries) or the NA98 LM (65k, 4.9M, 14.5M entries). Having said this, a pessimistic point of view for this simple n-gram approach is that even relatively large sized models, such as BN96 LM, are still not large enough to accommodate sufficient number of bi/trigrams.

When using the H4-train LM, the recall score for subclass <organisation> (.63) was relatively higher than <person> (.37) and <location> (.42), since there were more cues around <organisation> names[4] than the other two (although this statement is by observation without any statistical backing); as a consequence, bigrams and trigrams were more likely to be present in the LM. Furthermore, even without any cues, many <organisation> names contained multiple words, resulting in sufficiently high probability scores. Similar observations for number expressions (<money> and <percentage>), which usually occurred as longer sequences of words (often with other cues).

A secondary cause of inaccurate NE identification were errors in the BN and NA News training data produced by the automatic tagger[5]. Occasionally it also marked corpora with <name> tags when unresolvable type ambiguity occurred between <organisation>, <person>, and <location>. This inaccuracy seemed to contribute some of failures, for <organisation> in particular, when using the BN96 and the NA98 LMs.

Precision scores. NE annotation using the BN96 and the NA98 LMs achieved about the same level of precision as one using the H4-train model (except for temporal expressions <date> and <time>). Although the automatic marking on the BN and the NA News contained errors, it was compensated by a more reliable estimate of model parameters due to an increase in corpus size of over two ordered of magnitude. Precision scores for <person> and <location> were over 90%.

Because of a specification conflict, as noted earlier, the BN96-tagged and the NA98-tagged transcripts were poorly matched to temporal expressions (a precision of just over .3 for <date> class and well below .2 for <time> class). On the other hand, their precision scores for number expressions were good.

3.4. 1998 Hub-4 Evaluation Results

The n-gram based NE annotation system was applied to the 1998 DARPA/NIST *Hub-4E* NE identification task. The evaluation data contained over 30,000 words; half of which were selected from broadcast news shows taken between October and November 1996, the remainder from June 1998. The **test data set** consisted of (a) manually verified transcriptions and (b) the 1998 SPRACH recogniser output with 21% WER [8].

[3] For example, "CLINTON" was successfully identified as <person> when it occurred as

"... PRESIDENT CLINTON ..."

because the bigram "PRESIDENT <person>CLINTON" existed in the H4-train LM. It failed for a sequence

"... DOES IT HELP CLINTON WELL ..."

because the unigram probability for "<person>CLINTON" was lower than the bigram "<UNKNOWN> WELL" with no name tag on each.

[4] For example,

"THE DOW JONES INDUSTRIAL AVERAGE"

was a commonly observed phrase in the transcripts and thus "DOW JONES" was easily identified. Even simpler sequence

"THE DOW"

was sufficient to convince that "DOW" was <organisation>. In general, "THE" seemed a good clue to identify <organisation> names. Other such example was "THE WHITE HOUSE".

[5] One example was that the *LaSIE-II* system missed to identify "REPUBLICANS" as <organisation>.

site	transcript	*R*	*P*	*P&R*	*SER*
SPRACH	hand	.83	.84	.83	29.1%
	SPRACH	.72	.74	.73	47.1%
BBN	hand	.90	.91	.91	15.7%
MITRE	hand	.87	.90	.88	20.3%

Table 2: NE identification results on the 1998 *Hub-4E* evaluation data set. Scores for hand transcriptions and the 1998 SPRACH recogniser output (with 21% WER) are shown for our system; scores by BBN and MITRE are also reproduced for hand transcriptions. Numbers here were taken from the official NIST web site (`ftp://jaguar.ncsl.nist.gov/csr98/`). *R*, *P*, *P&R*, and *SER* denote recall, precision, precision&recall scores, and slot error rate [9] respectively.

Table 2 shows NE identification scores on this task. Our approach resulted in average precision and recall scores of around 83% on hand transcribed data, and 73% on the SPRACH recogniser output. For comparison, we also reproduce results (on hand transcriptions) reported by BBN and MITRE in this evaluation, both of which use an alternative statistical state machine approach [10, 11].

4. DISCUSSION

The n-gram approach presented in this paper resulted in precision and recall scores that were 5–10% worse than those reported by BBN and MITRE, even though those systems were trained only on the one million word H4-train annotated data. Ignoring technicalities, their methods both 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 are heavily dependent on existence of higher order n-grams in the model. The most straightforward way to improve the direct n-gram approach seems to be via the incorporation of constraints on a class level.

Acknowledgements. The authors wish to thank Kevin Humphreys, Hamish Cunningham, and Rob Gaizauskas for assistance in using the *LaSIE-II* NE identification system. This work was funded by ESPRIT project 20077 (“SPRACH”) and UK EPSRC grant GR/M36717.

5. 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] N. Chinchor, P. Robinson, and E. Brown, *Hub-4 Named Entity Task Definition (version 4.8)*. SAIC, August 1998.

[5] 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.

[6] N. Chinchor, *MUC-7 Named Entity Task Definition (version 3.5)*. SAIC, September 1997.

[7] 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.

[8] 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.

[9] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, “Performance measures for information extraction,” in *Proceedings of DARPA Broadcast News Workshop*, (Herndon, VA), February 1999.

[10] 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.

[11] 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.