BASELINE IE-NE EXPERIMENTS USING THE SPRACH/LASIE SYSTEM

Steve Renals *Yoshihiko Gotoh* *Robert Gaizauskas* *Mark Stevenson*

University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello St., Sheffield S1 4DP, UK

ABSTRACT

We have developed two conceptually different systems that are able to identify named entities from spoken audio. One (referred to as SPRACH-S) has a stochastic finite state machine structure for use with an acoustic model that identifies both words and named entities from speech data. The other (referred to as SPRACH-R) is a rule-based system which uses matching against stored name lists, part-of-speech tagging, and light phrasal parsing with specialised named entity grammars. We provide an overview of the two approaches and present results on the *Hub-4E IE-NE* evaluation task.

1. INTRODUCTION

This paper describes our participation in the *Hub-4E IE-NE* spoke. The SPRACH/*LaSIE* system for named entity (NE) identification in Broadcast News consists of two baseline systems:

SPRACH-S: a statistical system based on the NE tagged language modelling approach [1], which was originally introduced to enable name category information to be used in the construction of language models for very large vocabulary speech recognisers;

SPRACH-R: a rule-based approach [2, 3], ported from the *LaSIE* system used for text-based NE identification.

The stochastic finite state model approach is based on explicit word-level *n*-gram relations. We present an overview of the statistical system and a procedure for NE annotation. Then we describe key features for the rule-based approach. Comparison is made between the original text- and speech-based systems. The SPRACH-S and R systems were employed in the 1998 *Hub-4E IE-NE* evaluation. We report our results using the five sets of transcripts: a reference transcription, a transcription produced by the 1998 SPRACH recogniser used in the transcription evaluation (21% WER) and the three baseline transcriptions provided by NIST.

2. THE STATISTICAL SYSTEM: SPRACH-S

The SPRACH-S system consists of an NE tagged language model and a recently developed statistical NE tagger. A formal description of the NE tagged LM is provided in [1]. Technical details for the development and for the annotation procedure are presented in [4].

2.1. Named Entity Tagged LM

The basic idea of the NE tagged language model (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. Unigram extensions for less frequent names are attached in order to increase the overall vocabulary size.

For the evaluation, three NE tagged trigram LMs were estimated, each with an independent vocabulary set plus unigram extensions:

H4-train LM: derived from transcripts of the *Hub-4E* acoustic training data (approximately one million words with manual NE annotations) consisting of an 18k trigram vocabulary (*i.e.*, tag-word tokens), with a further 4k vocabulary in unigram extensions;

BN96 LM: estimated from 1996 BN text corpus for training/test data (150 million words with automatic NE annotations), consisting of a 65k trigram vocabulary, with a further 85k vocabulary in unigram extensions;

NA98 LM: estimated from a part of the 1998 North American News (NA News) corpus (1997-98 LA Times/Washington Post, 1996-98 Associated Press; 133 million words with automatic NE annotations), consisting of a 65k trigram vocabulary, with a further 145k vocabulary in unigram extensions.

Manual NE annotations were provided by MITRE and BBN (through NIST) and they conformed with the *Hub-4E* NE task specification. Automatic annotations were achieved using the *LaSIE-II* system [2]. Because the *LaSIE-II* was developed according to the *MUC-7* NE task specification, relative time expressions were also tagged for the BN and NA News corpora, conflicting with the *Hub-4E* specification.

2.2. NE Identification

After several trial runs using the development set (described later), 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. 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 were ignored.

The initial marking on speech transcripts was done using the trigram constraints 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 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).

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).

3. THE RULE-BASED SYSTEM: SPRACH-R

The SPRACH-R system described in this section was specifically developed for the 1998 *Hub-4E IE-NE* spoke. It uses a restricted and slightly modified version of the NE annotation component of the Sheffield *LaSIE-II* information extraction system, as entered in *MUC-7* [2] and described in detail in [3].

3.1. Key Features

The rule-based approach relies on: finite state matching against lists of single or multi-word names and NE cue words, part-of-speech tagging, and specialised NE parsing based on phrasal grammars for the NE classes. The key stages of processing are as follows:

Pseudo sentence segmenter. Since part-of-speech and parsing components of the system require text units of reasonable length (ideally less than 40 words; anything over 100 words becomes excessively slow), a trivial text segmenter breaks the text into "pseudo sentences" by breaking before certain closed class words (determiners, nominal pronouns, certain prepositions). The aim is not to find true sentence boundaries but to produce sensible length text chunks which are not broken in the middle of named entities[1].

Gazetteer lookup. Lists of single and multi-word names and name cues are used to tag the input. These lists include male and female first names, person titles (*e.g.*, "Mr.", "Mayor"), well-known locations and organisations, location cue words (*e.g.*, "Bay", "Harbour") and company designators (*e.g.*, "Corporation"). Case-insensitive finite state matching is carried out. Multiple tags may be assigned per word or multi-word.

Part-of-speech tagging. A version of the Brill transformation based part-of-speech tagger retrained for all upper case text is used to assign one of the Penn Treebank word classes to each word in the input.

[1] Our current work uses speech recognition language models that include sentence boundaries.

NE parsing. A bottom-up partial chart parser applies a set of regular NE grammars (one for each NE class, plus a general NE grammar and a default NE grammar). A typical rule has a form such as:

PERSON_NE ⟶ PERSON_FIRST_NAME PROPER_NAME

where PERSON_FIRST_NAME is a tag assigned by the gazetteer lookup stage, and PROPER_NAME is a tag assigned by the part-of-speech tagger. Since the results of parsing may be ambiguous (the same word sequence may be assigned multiple NE tags; overlapping word sequences may be assigned distinct NE tags), a "best-parse" algorithm selects unique, non-conflicting interpretations. This algorithm attempts to maximise lexical coverage while the number of distinct named entities found.

3.2. Annotating Speech Transcriptions

The SPRACH-R system was derived from one designed to do full information extraction on well punctuated, mixed-case newswire text. In addition to pseudo sentence segmenting and retraining the part-of-speech tagger on upper-case only text, there are a number of other differences between this system and the original *LaSIE-II* system:

- Parsing in SPRACH-R stops with NE parsing. In *LaSIE-II*, full sentence parsing is attempted using various phrasal grammars (NP, VP, PP, relative clause and sentence grammars). This impacts NE annotation since interpretations in which words are linked into larger phrases outside of an NE phrase would be preferred if such phrases were being sought and found, as they are in *LaSIE-II*.
- *LaSIE-II* attempts name matching across a text — *i.e.*, it attempts to match variant forms of a name (*e.g.*, "Bill Smith" and "Smith"). This can increase the accuracy of name classification significantly, but relies on texts being coherent in a way that typically the unsegmented speech transcriptions are not.
- *LaSIE-II* uses coreference to help in name recognition — *e.g.*, if "Ford" and "The company" can be co-referred then "Ford" may be accurately classified. Again, this assumes texts which are single, coherent stories.
- The relative time expression grammar developed for the *MUC-7* NE task was decoupled to conform to the *Hub-4E* NE task specification according to which relative time expressions are not to be tagged.
- Some of the person name lists used in the gazetteer lookup stage were modified to reduce ambiguities caused by single case text (*e.g.*, names such as "WILL", "MAY", and "ARE" were removed).
- Some of the NE grammars were modified slightly to prevent unclassified named entities from being passed forward — *LaSIE-II* allowed a category of unclassified named entities to be created during parsing and then further specified during name matching or coreference resolution.

4. Hub-4E IE-NE EVALUATION

We participated in the evaluation using the statistical and the rule-based NE annotation systems. For a **development data set** (1997 *Hub-4E* evaluation data) we report results of experiments using the manually verified reference transcriptions and the transcriptions from the 1997 CU-CON system (27% WER). A **test data set** (1998 evaluation data) consisted of reference transcriptions, the 1998 SPRACH recogniser output with 21% WER, and three baseline recogniser outputs.

LM	reference			CU-CON		
	R	*P*	*F*	*R*	*P*	*F*
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: SPRACH-S NE identification scores on the development set (1997 evaluation data). Results are shown for each of three component LMs (defined in the text) along with the merged system ("all"). *R*, *P*, and *F* denote recall, precision, and the F-measure.

	97 development set			98 test set (1)			98 test set (2)		
	R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
ref	.78	.84	.80	.80	.84	.82	.84	.85	.85
b1	-	-	-	.72	.76	.74	.77	.78	.77
b2	-	-	-	.73	.76	.74	.79	.79	.79
b3	-	-	-	.62	.66	.64	.67	.69	.68
sp	.66	.71	.68	.68	.73	.71	.74	.75	.75

Table 2: SPRACH-S 1998 *Hub-4E IE-NE* evaluation results using reference (ref), baseline (b1, b2, b3) and SPRACH (sp) transcriptions.

4.1. SPRACH-S

System Development. 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 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 an improvement in recall and F-measure.

In the following, we analyse NE annotation errors by closer inspection to the mark-ups on the development data set. [4] provides further description of errors using graphs and examples found in the annotated transcriptions.

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.*, an 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 tag) followed by some other word, rather than the unigram of a tagged word. 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).

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 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.

A secondary cause of inaccurate NE identification were errors in the BN and NA News training data produced by the automatic tagger. 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. Except for temporal expressions, NE annotation using the BN96 and the NA98 LMs achieved about the same level of precision as one using the H4-train model. Especially, precision scores for <person>, <location>, <money>, and <percentage> were easily over 90% for the former. Although the automatic marking contained some errors, it was compensated by a more reliable estimate of model parameters due to an increase in corpus size. Because of a specification conflict, the BN96-tagged and the NA98-tagged transcripts were poorly matched to temporal expressions (a precision of just over .3 for <date> and well below .2 for <time>).

1998 Hub-4E Evaluation Results. Table 2 shows the NE identification results for the SPRACH-S system on the 1998 evaluation data. 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.

4.2. SPRACH-R

The results of the rule-based approach are shown in Table 3, for both the reference transcriptions and the SPRACH and baseline recogniser transcriptions. Breakdown of the results by NE category for both 1998 test sets is shown in Table 4. Note that due to a porting bug no time, money or percentage NEs were identified by the SPRACH-R system.

On the reference transcriptions, the rule-based system returns an overall F-measure that 20-25% lower than those returned in *MUC-6* and *MUC-7*. The errors committed by the system may be divided into three classes:

	97 development set			98 test set (1)			98 test set (2)		
	R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
ref	.58	.86	.69	.64	.87	.74	.58	.83	.68
ref*	-	-	-	.66	.86	.75	.60	.82	.69
b1	-	-	-	.56	.81	.67	.53	.78	.63
b2	-	-	-	.56	.80	.66	.53	.78	.63
b3	-	-	-	.49	.75	.59	.45	.72	.55
sp	.48	.75	.59	.53	.78	.63	.49	.74	.59

Table 3: SPRACH-R 1998 *Hub-4E IE-NE* evaluation results using the reference (ref), baseline (b1, b2, b3) and SPRACH (sp) transcriptions. The ref* row shows post-evaluation results obtained after fixing matching bugs that affected time, money and percent classes.

class	number	*R*	*P*	*F*
<organisation>	423	.45	.71	.55
<person>	434	.56	.93	.70
<location>	712	.87	.90	.89
<date>	79	.53	.95	.68
<time>	19	0 (.50)	0 (1.0)	0 (.67)
<money>	79	0 (.70)	0 (.96)	0 (.81)
<percentage>	25	0 (.76)	0 (1.0)	0 (.86)

Table 4: SPRACH-R results by NE class using the combined 1998 reference transcriptions (test sets (1) and (2)). The numbers in parentheses for the time, money and percent tags are the values obtained post-evaluation, after fixing some matching bugs.

1. **Porting Errors.** These are the least interesting errors, and arise from rapid porting and insufficient testing of the *MUC-7* rule-based system to meet *Hub-4E* evaluation deadlines. They include errors such as failing to account fully for differences in protocols between transcriptions and newswire text for writing acronyms (*e.g.*, "C. N. N." and "CNN") and a trivial matching bug that caused all currency units (*e.g.*, "DOLLARS") to be missed in all transcriptions. These can be easily fixed and tell us little about the strengths and weaknesses of the underlying approach.

2. **Genre-Related Errors.** These are errors that arise because of differences in genre between the *MUC-7* newswire texts for which the rule-based system was developed and the broadcast news transcripts which *Hub-4E* addressed. We have noticed at least two such significant differences.

   First, newswire stories provide clearly delineated discourses in which a limited set of entities is introduced and then referred to in various ways (*e.g.*, "Winston Scott", later just "Scott"; "Bloomberg News Service", later just "Bloomberg"). The broadcast news transcriptions are not segmented into discourse units in any clearly identifiable way. Without a notion of "story boundary", techniques developed for matching variable forms of names across one story in newswire texts could not be used. Since the initial form of reference is usually fuller and hence more easily classified, inability to resolve subsequent references with earlier ones lead to a significant drop in recall (but attempts to do name matching across arbitrarily segmented portions of the transcriptions lead to even more exaggerated drops in precision).

   Second, the use of company designators (*e.g.*, "Inc.", "Ltd.") and personal titles (*e.g.*, "Mr.", "Dr.") appears to be much more limited in spoken news. These terms provide significant clues in text-based news stories, of which the rule-based system takes considerable advantage.

3. **Modality-Related Errors.** These are errors that arise because of the greater intrinsic difficulty of processing speech transcriptions over newswire text, *i.e.*, because of the loss of information in single case, unpunctuated text. For example, in mixed-case text a surname which is also a common noun (*e.g.*, "Butler") is easily recognised as a proper name, and the problem reduces to assigning it the correct name class; in all uppercase text this information is not available and other information must be used instead.

5. CONCLUSIONS

These experiments were preliminary experiments using baseline statistical and rule-based systems. The statistical system does not specifically model extent and has a context limited by the history of a trigram language model. It is also dependent on annotated training data, which we expanded using the existing *LaSIE-II* system. However, the statistical system is a very close match to the language model currently used in LVCSR systems, and it is straightforward to see how the NE tagged LM could be integrated into an LVCSR system. The rule-based system — which has produced good performance in previous *MUC* evaluations — was minimally modified to spoken rather than textual data and was not modified for the broadcast news domain. Although both systems are still under development, we are in a good position to investigate differences between statistical and rule-based approaches for information extraction. We also hope to investigate the possibility of constructing a hybrid system.

Acknowledgments. The authors would like to thank Kevin Humphreys for assistance in modifying the *LaSIE-II* NE identifier to work with spoken language transcripts. This work was funded by ESPRIT Long Term Research Project 20077 (SPRACH) and by UK EPSRC grants GR/K25267 and GR/M36717.

References

1. 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.
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 7th Message Understanding Conference (MUC-7)*, 1998.
3. T. Wakao, R. Gaizauskas, and Y. Wilks, "Evaluation of an algorithm for the recognition and classification of proper names," in *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, (Copenhagen), pp. 418–423, 1996.
4. Y. Gotoh and S. Renals, "Statistical annotation of named entities in spoken audio," in *Proceedings of the European Speech Communication Association (ESCA) Workshop: Accessing Information in Spoken Audio*, (Cambridge), April 1999.