Short Queries, Natural Language and Spoken Document Retrieval: Experiments at Glasgow University

Fabio Crestani[*], Mark Sanderson[†], Marcos Theophylactou,
Mounia Lalmas
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, Scotland

Abstract

This paper contains a description of the methodology and results of the three TREC submissions made by the Glasgow IR group (glair). In addition to submitting to the ad hoc task, submissions were also made to NLP track and to the SDR speech 'pre-track'. Results from our submissions reveal that some of our approaches have performed poorly (i.e. ad hoc and NLP track), but we have also had success particularly in the speech track through use of transcript merging. We also highlight and discuss a seemingly unusual result where retrieval based on the very short versions of the TREC ad hoc queries produced better retrieval effectiveness than retrieval based on more 'normal' length queries.

1 Introduction

This paper contains a description of the methodology and results of the ad hoc, NLP, and SDR submissions made by the Glasgow IR group (**glair**) to this year's TREC. The only common factor between the submissions is their

[*]Supported by a "Marie Curie" Research Fellowship from the European Union.
[†]Supported by VIPIR project of the University of Glasgow

use of a Glasgow built retrieval system, SIRE and this is introduced first in the paper. As the submissions are quite independent of each other, the rest of the paper is structured as an amalgam of three sub papers each with their own introduction, methodology, results and conclusions. The order of these sub papers is first, the ad hoc submission, second the NLP track, and finally the SDR track submission.

2 The SIRE Information Retrieval system

The system used in in the context of the work reported in this paper is a retrieval toolkit called *SIRE* (System for Information Retrieval Experimentation) developed "in-house" at Glasgow University by Mark Sanderson. SIRE is a collection of small independent modules, each conducting one part of the indexing, retrieval and evaluation tasks required for classic retrieval experimentation. The modules are linked in a pipeline architecture communicating through a common token based language. SIRE was initially used in research examining the relationship between word sense ambiguity, disambiguation, and retrieval effectiveness [8]. It proved to be a flexible tool as it not only provided retrieval functionality but a number of its core modules were used to build a word sense disambiguator as well. It was also used in the experiments for the Glasgow IR group submissions to TREC-4 and TREC-5 and is currently being used in a number of research efforts within the group.

SIRE is implemented on the UNIX operating system which, with its scripting and pre-emptive multi-tasking is eminently suitable for handling the modular nature of SIRE.

SIRE was chosen as the IR platform for the experiments reported in this paper because it implemented a probabilistic IR model we are very familiar with, based on the "TF-IDF" weighting schema [12]. Moreover, it was relatively easy to modify the code to take into account the characteristics of the new data.

A detailed description of the functionalities of SIRE is outside the scope of this paper. The system is currently public available for research purposes. The interested reader should contact Mark Sanderson for a copy of a short unpublished paper describing the system [7] and for the location of SIRE's binary files. The system has been successfully used by many students of the Advanced Information Systems M.Sc. of Glasgow University for their practical work.

3 Main ad hoc task: short queries and semi-automatic query expansion

In the ad hoc task of TREC the Glasgow IR group submitted three runs: **glair61**, **glair62**, and **glair64**. The main aim of this work was to investigate a means of improving retrievals for the very short queries of TREC-6. Because of their length, it was assumed that their use would result in poor retrieval and it would be necessary to expand them in some manner. The first two submissions (**glair61** & **glair62**) were aimed at testing such an expansion technique based on the manual identification of the senses of query words and the subsequent automatic expansion of those senses.

This work was somewhat overshadowed by the effectiveness results returned from the **glair64** submission - retrieval based on normal length queries (i.e. TREC query description fields) - which proved to be worse than the **glair61** results - retrieval based on the very short queries (i.e. their title fields). In other words, the very short queries were better than the normal length queries.

The rest of this section will first, describe the implementation, and results of the semi-automatic query expansion experiments and second, explore possible reasons for the drop in retrieval effectiveness found to occur when using the longer, and presumably more detailed, versions of the TREC queries.

3.1 Semi-automatic query expansion

A new feature of TREC this year was the introduction of the very short query task: ad hoc retrieval based on the title section of TREC queries. These queries were intended to mimic the type of query normally submitted to interactive IR systems by untrained, casual users. Their generation was governed by a set of guidelines[9], an extract of which is shown below.

> ... we would like you to make an effort in ensuring that the length of the titles is kept as short as possible. Please try to keep the length of the title to between 1 and 3 non-stop words. Only in exceptional circumstances would they be any longer, for example, if the title were some well known phrase or a long proper name. Do not worry if the title is not an accurate expression of the information need, this is a common feature of very short

queries: there is only so much that can be expressed in such a small number of words.

The very short queries generated from these guidelines were on average 2.5 non-stop words in length, as opposed to the normal length queries (based on the description field) which were 8.5 non-stop words in length. Figure 1 shows a couple of these queries (numbers 310 & 349) to illustrate these two query types.

```
<title> Radio Waves and Brain Cancer

<desc> Description:
Evidence that radio waves from radio towers or car phones
affect brain cancer occurrence.


<title> Metabolism

<desc> Description:
Document will discuss the chemical reactions necessary to
keep living cells healthy and/or producing energy.
```

Figure 1: Queries 310 & 349

It would probably be fair to say that there was an assumption among many involved in the decision to include these queries in TREC-6 that the effectiveness of any IR system retrieving from them would be poor when compared to retrievals using the more normal TREC queries based on the description field. With this preconception in mind, it was decided (by one of the authors) to explore the possibility of incorporating some type of query expansion into the very short queries. The one chosen was a semi-automatic form that required the manual identification of the sense of each query word followed by the automatic expansion of the identified senses with synonyms taken from a thesaurus. Similar ideas of mixing manual tagging with thesaurus based expansion have been reported by [13]. One of the conclusions drawn from this research was that expansion of shorter queries was more likely to improve retrieval effectiveness than expansion of longer queries. It was hoped that this situation would be encountered in the experiments on the very short queries of TREC. However, another conclusion of [13] was that use of automatic expansion methods could make queries decidedly worse. It was hoped that trying different forms of expansion in our experiments could counter these potential problems.

3.1.1 Implementation of experiments

There were three main components to this experiment: the document collection used, the retrieval system employed; and the thesaurus chosen to provide the sense definitions and synonyms. The collection was the 'A' collection as defined in the TREC-6 guidelines. The retrieval system employed was SIRE using standard IR features such as stop word removal, stemming and a $tf \times idf$ weighting scheme. The thesaurus used was WordNet [5], chosen because of it's coverage, ease of use and availability.

The first part of the expansion process involved the manual identification of query word senses. This was undertaken by one of the authors who looked up each query word in WordNet and assigned the sense closest to that word (this also involved the identification of the grammatical form that each word was used in). As WordNet stores phrases as well as words (e.g. 'land mine'), any possible query phrases were looked up before individual words were. Expansion of the word senses was simply a process of adding to the query exact synonyms of the senses. WordNet is quite sparing in its provision of synonyms, consequently queries were only expanded by a few words.

In choosing the precise form of expansion strategy employed for the TREC submission, experiments were run using the titles of the previous year's TREC queries (i.e. 251–300) on the 'B' collection of TREC-5. Results from these queries were disappointing: every expansion strategy tried was found to result in queries that produced lower retrieval effectiveness than that resulting from the unexpanded queries. Consequently, the 'least worst' strategy was chosen for submission in a vain hope that it would prove to be effective on the TREC-6 queries. The strategy consisted of expanding only the nouns of query words and leaving phrases unexpanded. In the experiments on queries 251–300, this strategy was found to improve 8 queries, leave 14 unchanged, and degrade 23 (the remaining 5 queries have no relevant documents). Unfortunately, this drop in effectiveness was repeated in the results returned from this year's TREC submission. The retrieval effectiveness of the queries after being expanded (**glair62**) was worse than the effectiveness of the unexpanded queries (**glair61**): with the expansion improving 3 queries, leaving 23 the same, and degrading 24.

As a footnote to this experiment, after submitting to TREC, some further expansion strategies were attempted on the 251–300 queries and a strategy was found that improved upon previous strategies, though still caused a drop in effectiveness, albeit a small one. The strategy was motivated from work reported by [8] which showed how the frequency of occurrence of the senses of words was skewed so that the most common sense of a word typically accounted for the majority of occurrences of that word. With this information in mind, it was surmised that query words used in their commonest sense did not need expansion as their sense would be so prevalent in the collection, expansion terms would more likely introduce error than help retrieve documents containing this sense. If, however, a query word was used in one of it less common senses, expansion might be useful in ensuring that documents containing that sense was retrieved. Using this strategy of only expanding the less common senses of query words on the TREC queries 251–300 resulted in 4 queries being improved, 36 unchanged, and 5 degraded. Information on the frequency of occurrence of word senses was gained from WordNet and not from the collection the experiment was conducted on. The increased number of unchanged queries is not surprising given that fewer expansions took place.

3.1.2 Conclusions

The strategy of targeting query words using a less common sense may be a promising strategy, though obviously one that requires much improvement before it can be employed in any retrieval system. It has not yet been tested on the TREC-6 queries 301–350 and this is one of the future aims of this work.

3.2 Short vs long: small ones are more juicy?

As was stated in the introduction to this section, the results from the query expansion experiments were over shadowed somewhat by the results of the **glair64** submission showing that retrievals based on the description part of TREC queries were worse than retrievals based on the title sections. Contrary to expectations, it would appear that the compact queries of the title field are in general better than the more verbose queries of the description field.

3.2.1 Brief discussion

In this section a brief discussion of the possible reasons for these results are presented along with speculation on possible changes to query design in future TRECs.

Are long queries cursed? There is a well known result in retrieval research showing, in the context of relevance feedback at least, that there is an optimum size of query for producing the best retrieval effectiveness. This effect, sometimes called the 'curse of dimensionality'[12], has been shown to exist on a number of retrieval systems [2, 8, 3] including SIRE (the retrieval system employed in these experiments). Therefore, one explanation for the drop in effectiveness found in the **glair64** result could be due to this curse. Indeed, it does appear to be a factor. Figure 2 shows a scatter plot of average precision against query length for the 50 queries of TREC-6 (301–350), showing that at longer query lengths, average precision is generally lower. This trend, however, is not strong and other explanations should be examined before entirely blaming the result on the curse.

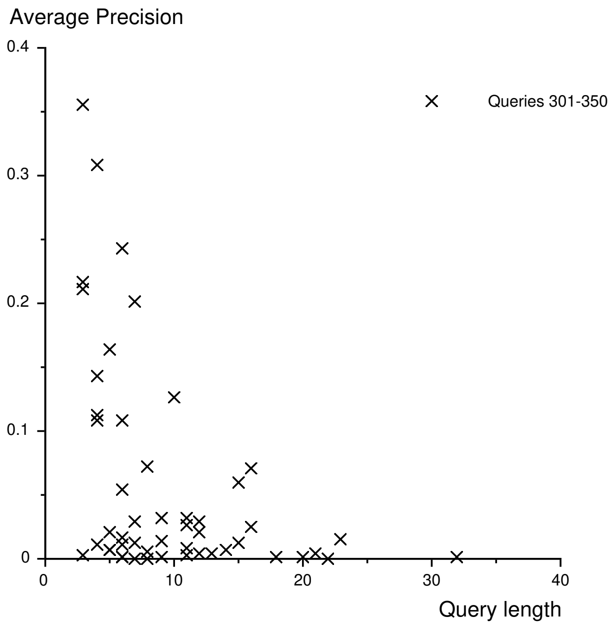

Figure 2: Scatter plot of average precision versus query length

Are the descriptions any good as queries? As can be seen in the two example queries in Figure 1, the description fields are written to be explanations of information need intended for human consumption. From the point of view of a retrieval system, they contain seemingly useless phrases such as 'document will discuss' (phrases that seasoned TREC participants have in their stop lists) and sometimes clarifications that of information need that would be hard for a retrieval system to detect. Unless a retrieval system can parse the natural language of a description field, such subtleties will be lost. With this in mind, it is questionable if comparisons between the title and description sections are entirely fair as the two fields were not created for the same purpose. Indeed, there are a few queries in this year's TREC where one sees the title and description being used in a complimentary manner. For example query 349 requesting documents on the processes of living cells: the description contains rather general and ambiguous words, where as the title field is the single word 'metabolism' (rather like a question and accompanying answer). The very short version of this query produces good retrieval, but the longer version (minus this highly descriptive word) performs much worse. Like the previous explanation, it is not suggested that this difference between the description and the title fields is the sole reason for the drop in effectiveness on the longer queries, but it would appear to be a factor.

In order to eliminate it, it might be necessary to alter the guidelines for generating the description field possibly making it less of an naturally expressed request for information, more a simple list of words. In addition, it would be necessary to ensure that the title and description fields are kept independent of each other to avoid the complimentary type of query shown in Figure 1.

4 The Natural Language Track

We have developed a document retrieval model that uses noun phrases and single word terms for indexing and the retrieval processes [11]. The model is based on the Dempster - Shafer (D-S) theory of evidence [10] which is a generalisation of the Bayesian approach. The experiments were carried out on the 'B' collection.

4.1 Brief overview of the Dempster-Shafer theory

The D-S theory is a theory of uncertainty that assigns *belief* to propositions. A particular characteristic of the theory is that the belief of a proposition, x, does not necessarily imply that the belief associated to the negation of the proposition is $1 - x$ (as happens in probability theory). In the absence of any other evidence to support the negation of the proposition, the remaining belief is assigned to the entire proposition set, and represents the *overall uncertainty* or *uncommitted belief*. The full understanding of the D-S theory is not the purpose of this paper. We only give the necessary information for the understanding of the document retrieval model developed.

The D-S theory uses a number in the range $[0, 1]$ to assign *exact* beliefs to mutually exclusive propositions of a *frame of discernment* Ω. The assignment is represented by a *basic probability assignment* usually denoted by m:

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{p \in \Omega} m(p) = 1$$

The belief values assigned must always sum to one. A belief assigned to Ω itself represents the uncommitted belief.

A fundamental function in the D-S framework is the belief function. The function calculates the total belief $\mathrm{Bel}(p)$ committed to the proposition p, from the available evidence (as expressed by the basic probability assignment):

$$\mathrm{Bel}(p) = \sum_{q \rightarrow p} m(q)$$

In contrast to the $m(p)$, which calculates the exact belief to p, $Bel(p)$ calculates the total belief committed to p.

4.2 Noun phrase extraction

We use a part of speech tagger module and a noun phrase extractor module for the extraction of noun phrases from the 'B' collection and TREC-6 queries 301–350. Tagging of all the text in document/query was performed followed by the extraction of several tag patterns considered to be noun phrases. Stop words were then deleted from noun phrases and the remaining words were stemmed using the Porter stemmer.

The Natural Language Processing modules used were designed and implemented at the Language Technology Group (LTG) of the Human Communication Research Centre (HCRC), University of Edinburgh. The tagger is a state-of-the-art tagger and is a resource used in the Knowledge Acquisition Workbench [4], currently under development. The tagger achieves 96-98% accuracy if all the words in the text are found in the taggers lexicon, and 88-92% if unknown words appear in the text.

4.3 Indexing and retrieval

4.3.1 Document indexing

Noun phrases extracted from documents were combined with single terms for the formation of a *frame of discernment* for the 'B' collection. For all the single terms of the document collection, all the 2^S boolean combinational elements were generated using the terms (S being their number), the negations ($\neg$) of these terms and the boolean conjunction ($\wedge$). These boolean elements represented the basic propositions of the constructed frame.

Suppose that a document collection contains only the two single terms "*red*" and "*wine*". We obtain the following four (basic) propositions in the frame Ω:

p_0	$\neg red \wedge \neg wine$
p_1	$\neg red \wedge \quad wine$
p_2	$red \wedge \neg wine$
p_3	$red \wedge \quad wine$

Any valid combination of the above four propositions (e.g., $p_1 \vee p_2$) is also a proposition of the frame Ω.

A basic probability assignment was associated with each document D_i. Its values were derived from the document frequency characteristics. The general weighting formula used in the first two runs (**Gla6DS1**, **Gla6DS2**) was:

$$m_i(p_j) = \begin{cases} \frac{\mathrm{FREQ}_i(x_j)}{\mathrm{TOTFREQ}_i} \cdot \log_N \frac{N}{n(x_j)} & j \neq 0 \text{ and } x_j \text{ is a term} \\ \frac{\mathrm{FREQ}_i(x_j)}{\mathrm{TOTFREQ}_i} \cdot \min\limits_{w \in x_j} \{\log_N \frac{N}{n(w)}\} & j \in 0 \text{ and } x_j \text{ is a noun phrase} \\ 1 - \sum\limits_{x_k \in D_i} m_i(x_k) & p_j = \Omega \\ 0 & j = 0 \end{cases}$$

where:

1. p_j is the disjunction of (basic) propositions in the frame for which is constructed upon the single term or noun phrase x_j where x_j holds true. $p_0 = \perp$ so $m_i(p_0) = 0$. $m_i(\Omega) = m_i(\top)$ represents the uncommitted

belief of document D_i (Ω can be viewed as the disjunction of all the basic propositions (except $\perp$), that is the true proposition $\top$.

2. $\mathrm{FREQ}_i(x_j)$ is the number of occurrences of x_j in document D_i.
3. $\mathrm{TOTFREQ}_i = \sum_{x_k \in D_i} \mathrm{FREQ}_i(x_k)$ is the number of total occurrences in document D_i.
4. $n(t_j)$ is the number of the documents in the collection that contain the term x_j.
5. $w \in x_j$ are all the single words in the noun phrase x_j.
6. $\log_N\left(\frac{N}{n(x_j)}\right)$ is the inverted document frequency (IDF) weight of the term x_j. We used the logarithm with base N so the IDF is in the interval $[0, 1]$.

The weighting schema used is version of the classic TF-IDF using normalised TF and normalised IDF. The TF factor is normalised with the length of the document ($\mathrm{TOTFREQ}_i$) and the IDF factor is normalised with the logarithm of N. The D-S restriction for total belief being always equal to one motivated the normalised TF and IDF factors. The IDF value of noun phrases is always equal to the minimum IDF value of the single terms that constitute the noun phrase.

For the third run (**Gla6DS3**) the TF factor used is different for single terms. For each single term appearing in a noun phrase the frequency assigned to it is only the number of its occurrences in the document as a stand alone term (without counting its occurrences when it appears in a noun phrase).

4.3.2 Queries and Retrieval

The queries used in the three runs fall in these two categories:

Single term queries: Only single terms are used. This category was used in the first (**Gla6DS1**) and the third run (**Gla6DS3**).

Noun phrase queries: The noun phrases are extracted from queries were considered. The single terms that appear only in a noun phrase and not as stand alone single terms in a query, are used in the query only as part of the extracted noun phrase. This category of queries was used in the second run (**Gla6DS3**).

Queries are mapped onto the frame of discernment as a proposition:

$$Q = \bigvee_{p_k \in \text{query}} p_k$$

p_k are the propositions for terms x_k as defined in the document representation. The disjunction ($\vee$) is used since it is difficult to derive from a natural language query whether a user wants to find documents about "*red wine*" or documents about "*red*" or "*wine*" unless the former is found as a noun phrase in the query. If the term x_k can not be expressed as a proposition in the frame Ω then p_k is assigned the empty proposition $\perp$.

For measuring relevance of a query to a document the belief function of the D-S theory was used. The relevance of a document to a query is formulated as:

$$\text{Bel}_i(Q) = \sum_{p \to Q} m_i(p)$$

In documents where the belief value is zero there is no relevance of the document to the query. None of its indexing proposition implies the query proposition. For a document collection, all the estimated relevant documents to the query ($\text{Bel}_i(Q) > 0$) can be ranked using the belief value of each document for ranking. For example, a query with only the word "*wine*" will have belief value equal to the basic probability assigned to the propositions built upon the word "*wine*" (these are the propositions p_1 and p_3 in the table). A query with the noun phrase "*red wine*" will have belief value equal to the basic probability assigned to the propositions derived from the two words "*wine*" and "*red*" (this is the proposition p_3 in the table).

4.4 Results

The results obtained cannot be considered successful. Though the theoretical framework supporting the model is sound, the application of the proposed basic probability assignments and the belief function seems to lower precision when belief is given to noun phrases.

The main reason is that words with low IDF values are also existent in many noun phrases. For example, in the 'B' collection, the word "*account*" is a very frequent term. When it appears in noun phrases the belief value of the stand alone word increases. If a query requests for "*swiss account*"(interpreted as as a disjunction), a document containing the noun phrase "*current account*" three times will be retrieved with high belief even though the word "*swiss*" is not contained in the document. This happens when the single word query approach is used (runs **Gla6DS1** and **Gla6DS3**).

A method for solving the above problems is to use the noun phrase queries (run **GlaDS2**). Unfortunately, this query approach retrieves only documents containing the noun phrase of the query. In the previous example the noun phrase "*current account*" will retrieve documents containing it but, it will not retrieve documents that have only the words "*swiss*" or "*account*" which are relevant to the query (though they do not contain the noun phrase "*swiss account*".

In brief, the main problem of the belief function as used in this model falls into two cases:

1. If single word queries are used it increases the belief of frequently unwanted terms in irrelevant documents, thus lowering dramatically precision.
2. If noun phrase queries are used the belief function is very specific in retrieval, and recall gets strongly affected.

Another major problem is the use of document length normalisation to the basic probability assignment which misleads the retrieval of short documents.

5 The Spoken Document Retrieval Track

5.1 The Abbot Speech Recognition System

The speech recognition system we used for the participation to the SDR track was kindly made available to us by the Speech and Hearing Research Group of the Department of Computing Science of the University of Sheffield track. We did not have to perform any speech recognition on the speech data, since we were given the transcripts by the Sheffield group. Nevertheless, we fell obliged to give a few details about the speech recognition system they used, referring back to their article at TREC-6 for more. The system they used is Abbot.

Abbot is a speaker independent continuous speech recognition system developed by the Connectionist Speech Group at Cambridge University and now jointly supported by Cambridge and Sheffield Universities with commercialisation by SoftSound.

The Abbot system grew out of a PhD on recurrent neural networks at the University of Cambridge. It was further developed under the ESPRIT project "Auditory Connectionist Techniques for Speech" and then the ESPRIT project "WERNICKE: A Neural Network Based, Speaker Independent, Large Vocabulary, Continuous Speech Recognition System". Currently further development is being funded by the Framework 4 projects "SPRACH: Speech Recognition algorithms for connectionist hybrids" and "THISL: Thematic Indexing of Spoken Language".

The system is designed to recognise British English and American English clearly spoken in a quiet acoustic environment. The system is based on a model that is a combination of a connectionist and a Hidden Markov model [6].

5.2 Experimenting Probabilistic Retrieval of Spoken Documents

In this section we report a brief account of the strategies we used for the two runs for the SDR track. A more detailed account of these techniques is reported in [1].

The PFT weighting schema

One of the characteristics of the data we had available from the Abbot speech recognition system is the uncertainty associated to each word recognised by Abbot. The following is an example of part of a srt file produced by Abbot.


```
<Episode Filename=a960521.sph Program="ABC_Nightline"
Scribe="obert_markoff" Dat e="960521:2330" Version=4 Version_Date=961011>
.
.
.
<Section S_time=0.000 E_time=61.320 Type=Filler ID="a960521.1" >
<Word S_time=1.76 E_time=2 Prob=-1.873> IT'S </Word>
<Word S_time=2 E_time=2.048 Prob=-0.9346> A </Word>
<Word S_time=2.048 E_time=2.656 Prob=2.025> QUESTION </Word>
<Word S_time=2.656 E_time=2.832 Prob=-0.6394> THAT </Word>
<Word S_time=2.832 E_time=2.992 Prob=-0.3682> WILL </Word>
<Word S_time=2.992 E_time=3.36 Prob=1.188> MAKE </Word>
<Word S_time=3.408 E_time=3.488 Prob=-0.9622> A </Word>
<Word S_time=3.488 E_time=3.872 Prob=2.335> LOT </Word>
```

```
<Word S_time=3.872 E_time=3.984 Prob=0.4647> OF </Word>
<Word S_time=3.984 E_time=4.672 Prob=5.322> AMERICANS </Word>
<Word S_time=4.672 E_time=4.864 Prob=-0.4521> THINK </Word>
<Word S_time=6.882 E_time=6.994 Prob=-2.392> TO </Word>
<Word S_time=6.994 E_time=7.234 Prob=-1.807> HAVE </Word>
<Word S_time=7.234 E_time=7.346 Prob=-3.124> TO </Word>
<Word S_time=7.91 E_time=8.086 Prob=-0.2239> YOU </Word>
<Word S_time=8.086 E_time=8.294 Prob=0.1139> SAY </Word>
<Word S_time=8.294 E_time=8.454 Prob=-2.961> TO </Word>
<Word S_time=8.454 E_time=8.95 Prob=-3.794> ONE </Word>
.
.
.
</Section >
```

These measures of uncertainty are incorrectly called probabilities, as an explanation of the way they are computed will clarify:

1. For a given time segment, the neural network at the heart of Abbot provides a set of posterior probabilities for each phoneme. These are the "acoustic probabilities".

2. To facilitate the decoding, the acoustic probabilities are converted into scaled likelihoods by dividing by the prior probability of the phoneme.

3. During decoding, a search is performed using the acoustic probabilities and the language model to find the most likely sequence of words for that utterance.

4. As each word is defined as a sequence of phonemes, the score for that word is obtained by summing the scores of the individual phones which constitute that word. (Summing because Abbot works with log probabilities).

Although they are not probabilities, we can still consider them as weights expressing the confidence given by Abbot in the correct recognition of words. This gave us the idea of combine these weights with the probabilistic model underlying SIRE.

The probabilistic model used by SIRE assigned to every index term extracted from the text of a document a weight that is a combination of two different discrimination measures: the IDF and the TF. The IDF of a term is a collection wide weight, since it is calculated taking into account the distribution of the term inside the whole collection. The TF of a term is instead a document wide weight, since it is calculated taking into account the distribution of a term within a document. The TF is of particular interest in our discussion. The TF of a term is usually calculated as a normalised sum of the number of occurrences of that term in the document. If the occurrence of a term is a binary event, then:

$$occ.(x_j) = \begin{cases} 1 & \text{if } x_j \text{ occurs in } d_i \\ 0 & \text{otherwise} \end{cases}$$

Therefore, in its simplest definition, the frequency of occurrence of a term is defined as follows:

$$freq_i(x_j) = \sum_{d_i} occ.(x_j)$$

We decided to use the probabilities Abbot assigns to words as a way of devising a more general definition of occurrence. We decided to use the following definition of occurrence:

$$occ'.(x_j) = \begin{cases} Prob(x_j) & \text{if } x_j \text{ occurs in } d_i \\ 0 & \text{otherwise} \end{cases}$$

Therefore the frequency of occurrence of a term is now defined as:

$$freq_i(x_j) = \sum_{d_i} Prob(x_j)$$

This definition of frequency is the one used to redefine TF as follows:

$$PTF_{ij} = freq_i(x_j)$$

We called *PFT* (Probabilistic Term Frequency) this new definition of TF.

The above definition is quite intuitive. While TF measures the importance of a term in the context of a document as a function of the number of occurrences of the term, PTF weights the number of occurrences of a term with the confidence assigned every time to the recognition of the occurrence of the term. In fact, it is intuitive that the PTF of a term should be higher in the case the term being recognised as present in the document with high confidence values, that in the case of being recognised with low confidence values. In the latter case, in some instances, the term may have been mistaken for another term and may not even be present in the document.

In some of the experiments that follow we tried to see if a PTF-IDF weighting schema gives better performance than the classical TF-IDF. The actual formula for the PTF used in these experiments is, for reasons that we will not discuss here, the following:

$$PTF_{ij} = K + (1-K)\frac{freq_i(x_j)}{maxfreq_i}$$

Generating a weighting schema by merging different transcriptions

In the previous section we have taken advantage of a particular feature of the transcription we had available, the probabilities assigned by Abbot to words in the transcription. We used these probabilities to generate a new weighting schema for the words in the transcription. However, a few questions that we posed ourself were: are these probabilities reliable? Is there any other strategy that we could use to generate confidence (or uncertainty) values to assign to recognised words?

Another, perhaps naive, strategy that we decided to test was again due to our particular situation. We had availability of two different speech recognition transcript for the same speech data. A first analysis of the two transcripts shows large differences in recognition. Here is a short example:

BSRT (NIST/IBM recogniser) :

```
<Section S_time=0.000 E_time=61.320 Type=Filler ID="a960523.1" >

I will talk about blacks and winds we eventually go wrong a
of the tough question who he hid ...

</Section>
```

Abbot (Sheffield recogniser) :

```
<Section S_time=0.000 E_time=61.320 Type=Filler ID="a960523.1" >

we talked about blanks and whites we eventually get around
```

```
to the tough question his own unions say well ....

</Section>
```

DTT (Actual transcript) :

```
<Section S_time=0.000 E_time=61.320 Type=Filler ID="a960523.1" >

when we talk about blacks and whites we eventually get around
to the tough question some of you are ...

</Section>
```

It is easy to spot the errors made by the two speech recognition systems. One interesting fact is that there are many cases of words correctly recognised by one system and wrongly by the other. For example, the word "blacks" has been correctly recognised by BSRT and wrongly by Abbot, while the word "white" has been correctly recognised by Abbot and wrongly by BSRT. If one of these two words would have been used in a query, the IR system could not avoid retrieving only the document in which the word has been recognised correctly.

This suggested merging the two speech recognition transcripts. In this case the correct recognition of one system could compensate for the wrong ones of the other system. Moreover, using the classical TF-IDF weighting schema, if a word has been correctly recognised by both systems, then it will have a larger frequency of occurrences and this will increase its weight in the context of the document. On the other hand, a word that has been wrongly recognised by one of the speech recognition systems will have a small frequency of occurrence (unless it has been consistently recognised wrongly, a case that we suppose does not happen frequently) and therefore a lower weight in the context of the document. We called *Merged* this weighting schema.

5.3 Results

We will not discuss the figures returned from TREC in detail in this paper. We will just note that:

- the R1 run (**gla6R1**, using hand transcripts) is right on the median value;
- the B1 run (**gla6B1**, NIST/IBM data) is slightly above the median value;

- the S1 run (**gla6S1**, using the PTF strategy with Abbot data) is below the median value, clearly, if the PTF weighting scheme is to be of any use, it requires further work;
- the S2 run (**gla6S2**, using a merged NIST/Abbot collection) is above the median value and better than both the B1 run and the S1 run. In fact, under some of the evaluation measures listed in the results file (particularly the mean reciprocal) the S2 run is almost as good as the R1 run: the manual transcripts! In all the tests using merging, we found it to be always better than retrieval on the individual collections and we feel this provides some evidence towards regarding merging transcripts as a consistently good strategy in retrieval of spoken documents.

5.3.1 Conclusions and future works on SDR

This was our first experience in dealing with retrieval of spoken documents and we are pleased with the results of the initial efforts. Cross comparisons between groups with their alternate IR strategies and different recognisers is not easy. Our impression of the trend of results, however, is that no amount of clever retrieval strategies will compensate for a poorly recognised transcript. We certainly feel that our relative success in retrieving spoken documents has much to do with the quality of transcript generated by the Abbott System of Sheffield University.

6 Conclusions

To conclude, our participation to TREC-6 was a very interesting one and useful one in all three the tracks we took part in. The results achieved, that we only briefly reported in this paper but that are summarised at the end of this proceedings, encourage us to pursue our future participation for next TREC at least in the short queries and in the SDR tracks.

References

[1] F. Crestani, and M. Sanderson. Retrieval of Spoken Documents: First Experiences. Departmental Research Report TR-1997-34, Department of Computing Science, University of Glasgow, UK, October 1997.

[2] D. Harman. Relevance feedback revisited. In *Proceedings of ACM SIGIR*, pages 1–10, Copenhagen, Danmark, June 1992.

[3] M. Magennis and C.J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of ACM SIGIR*, pages 324–332, Philadelphia, PA, USA, July 1997.

[4] A. Mikheev and S. Finch. A workbench for finding structure in texts. In *Proceedings of the Applied Natural Language Processing (ANLP-97), Washington D.C.*, April 1997.

[5] G. A. Miller. A lexical database for english. *Communication of the ACM*, 38(11):39–41, 1995.

[6] T. Robinson and M. Hochberg and S. Renals. The use of recurrent networks in continuous speech recognition. In C. H. Lee and K. K. Paliwal and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, pages 233–258. Kluwer Academic Publishers, 1996.

[7] M. Sanderson. System for information retrieval experiments (SIRE). Unpublished paper, November 1996.

[8] M. Sanderson. *Word Sense Disambiguation and Information Retrieval.* PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, 1996.

[9] M. Sanderson and R. Wilkinson. Guidelines for generating very short TREC queries, 1997. `http://www.dcs.gla.ac.uk/~sanderso/guidelines.html`.

[10] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[11] Marcos Theophylactou. Document Retrieval using Natural Language Processing and the Dempster - Shafer Theory of Evidence. Master's thesis, University of Glasgow, Department of Computing Science, September 1997.

[12] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[13] E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of ACM SIGIR*, pages 61–69, Dublin, Ireland, July 1994.