様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　２９　日現在

機関番号：　５３３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３５０３５５

研究課題名（和文）統合知識を活用した自学自習用講義ビデオ検索システムの開発

研究課題名（英文）Development of self-study lecture video retrieval system using integrated knowledge

研究代表者

金寺　登（KANEDERA, Noboru）

石川工業高等専門学校・その他部局等・教授

研究者番号：５０１９４９３１

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：近年，講義ビデオなど音声情報を含むデータが蓄積され，一部は一般に公開されているが，情報量が増大するにつれて，必要な情報を検索することが困難になってきた。そこで，英語の講義ビデオにも対応した関連ビデオ検索システムを開発し公開した。このシステムでは，現在閲覧している講義ビデオシーンなど注目しているシーンによって関連する音声情報内容検索が行える。
検索方法として，シーンベクトルの線形補間がシーン間検索においても効果的であることを確認した。また，音声認識性能とシーン間音声情報検索性能の関係を明らかにした。

研究成果の概要（英文）：Along with increased multimedia information with spoken content, spoken document retrieval has attracted attention. Therefore, we developed a system that can register and retrieve videos, including lecture videos. This system extracts spoken content from videos and converts it to text by automatic speech recognition, this text can be targeted by queries. This system can also retrieve video scenes similar to the scenes currently being viewed.
We confirmed the linear interpolation is effective in related spoken document retrieval, where the linear interpolation use first subtopic and whale subtopic in the each video. We also reveal the relationship between speech recognition performance and related spoken document retrieval performance.

研究分野：　総合領域

キーワード：　教材情報システム　音声情報内容検索　シーンベクトルの線形補間　音声認識　講義ビデオ

１．研究開始当初の背景

ネットワーク上に蓄積された情報があまりに多いため，通常の検索では必要な情報が不要な多くの情報に埋もれてしまう。特にビデオ情報は，音声認識等でテキスト化する際に誤りが混入するため，通常のテキスト検索方法だけでは，十分な検索性能が得られない。

そこで，NTCIR (NII Testbeds and Community for Information access Research) では，情報検索・テキスト要約・情報抽出・質問応答など情報アクセス技術の研究をより発展させることを目的としたNTCIR 評価ワークショップが開催されている。この NTCIR ワークショップ評価タスクの一つである SpokenDoc-2 では，実際の検索環境に近い条件(自由発話音声を対象、未知語を含む検索課題) における共通タスクを設定し，大規模な音声ドキュメント検索タスクの評価が実施されている。我々もこの評価タスクに参加し，知識を活用したキーワード補完，コンテンツ補完による検索方法を提案し，検索性能が高いとの評価を得た。また，サブワード検索による音声認識誤りへの対応効果を確認した。さらにこれらの技術を実装したシステムを開発し公開した(http://sail.i.ishikawa-nct. ac.jp/)。これにより，組織内に蓄積したビデオ情報を効率的に検索できるようになった。

２．研究の目的

学習したい内容に関連する世界中のビデオ教材や各種情報を検索するためには，検索候補がさらに広がるため，検索精度の低下が大きな問題となる。検索精度を維持するには，これまで以上の知識の活用が必要となる。利用可能な知識には，WordNet, 日本語 WordNet, WikiPedia,一般的な辞書などがある。WordNet を用いると語の上位概念・下位概念を知ることができる。日本語 WordNet は WordNet を日本語化したものである。キーワードの下位概念が情報検索に有効であることがわかっている(N.Kanedera, 2013)。WikiPedia には，語の説明,関連項目,カテゴリーなどの情報が収められている。これらの内，語の説明と関連項目を併用して検索キーワードを拡張すると検索性能が向上することを予備実験により確認している。

より精度の高い情報検索を実現するためには，これまでの各種知識源を個別に利用するのではなく，有機的に統合しなければならない。具体的には辞書，翻訳技術や多言語に渡る知識ネットワークの構築と活用が必須である。

本研究では，これまでの研究成果である知識を活用したビデオ検索技術に，多言語に渡る知識ネットワーク情報を組み込み，人間と同等以上に関連講義ビデオシーンを検索できるシステムを開発することを目的とする。本システムが開発されれば多くの時間と労力を要した検索作業が即座に自動的に行われる。

３．研究の方法

平成 26 年度は，国内の複数の教育機関のビデオ教材を自動収集し，システム性能を評価した。まず各種の知識情報を活用し，複数の教育機関のビデオ教材各シーン間の関連を自動収集した。自動収集結果と人間による収集結果を精度と速度の面で比較した。

平成 27 年度は，自動翻訳技術を利用し，海外のビデオ教材各シーンと国内のビデオ教材各シーン間の関連を自動収集した。海外及び国内のビデオ教材関連シーンの自動収集結果と人間による収集結果を精度と速度の面で比較した。

平成 28 年度は，関連講義ビデオ自動収集検索システムの開発・公開した。

(1) シーン間音声情報内容検索環境

① シーン間音声情報内容検索評価データ

各種パラメータの値は，NITCIR9 を用いて最適化する。注目しているシーンをキーとした関連音声内容検索方法の評価には，NITCIR10 を使用する。

NITCIR9 の検索対象は日本語話し言葉コーパス CSJ に含まれる学会講演及び模擬講演 2702 講演であり，30 文毎に分割したシーンの検索を行う。NTCIR-9 で提供されたドライラン 39 クエリを使用し，各種パラメータを最適化する。また，NTCIR-9 で提供された単語単位の音声認識テキストを利用する。

シーン間の検索性能を調査するために，NTCIR-10 で提供された 120 クエリの正解区間データを利用する。NITCIR10 の検索対象は音声ドキュメント処理ワークショップのコーパス(SDPWS)の 104 講演であり，30 文ごとに分割すると 1408 シーンとなる。

図１のように各クエリに対して複数の正解区間がある。そこで，正解区間の一つを取り出し，対応する関連区間として，取り出した正解区間以外の正解区間を用いることとした. その結果, 1408 クエリ区間に対して，33204 の正解区間が得られた。このようにして得られたシーン間音声情報内容検索評価データ例を図２に示す。図２では，講演番号 07-24 の 5 文目から 8 分目のシーンに対応する音声認識されたテキストが検索クエリとなり，検索クエリの後に記載された講演番号とその対応区間が正解区間となる。

以上の方法は，通常の音声内容検索コーパスがあれば，シーン間検索コーパスを自動生成できるため，大変有用と考えている。

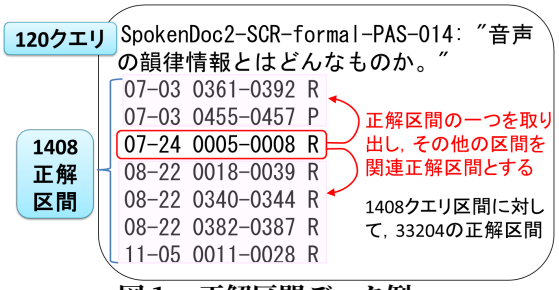

図１　正解区間データ例

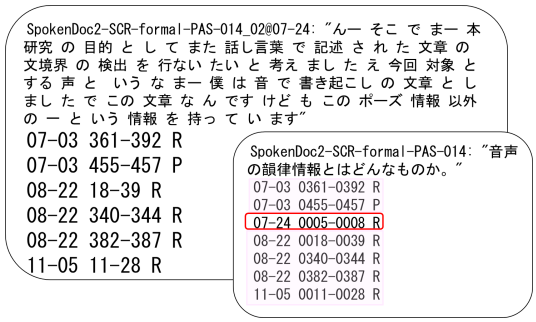

図２　シーン間音声情報内容検索データ例

シーン間音声情報内容検索に用いる音声認識テキストには，NTCIR-10 で提供された単語単位の「REF-WORD-MATCHED」と「REF-WORD-UNMATCHED」の２種類を利用した。「REF-WORD-MATCHED」は，CSJ から学習された単語単位の trigram 言語モデルを用いて音声認識されたテキストであり，SDPWS コーパスに対して，単語正解率 68.4%，単語正解精度 63.1%である。「REF-WORD-UNMATCHED」は，新聞記事から学習された単語単位の trigram 言語モデルを用いて音声認識されたテキストであり，SDPWS コーパスに対して，単語正解率 48.4%，単語正解精度 43.7%である。

シーン間の検索性能調査では，検索元のシーン及び検索対象のシーン共に「REF-WORD- MATCHED」を用いた場合（以後 match と表記する），検索元のシーン及び検索対象のシーン共に「REF-WORD-UNMATCHED」を用いた場合（以後 unmatch と表記する）について検索実験を行う。

② 評価指標

評価尺度には各クエリの平均適合率をクエリ数で平均した MAP(Mean Average Precision)を用いる。クエリ q に対し，全正解文書数を $|R_q|$，検索された上位 1000 件に含まれる正解文書を $r_1, r_2, \cdots, r_M$ $(M \le |R_q|)$ としたとき，平均適合率 $AveP_q$ 及び MAP は，以下のように計算できる。

$$AveP_q = \frac{1}{|R_q|}\sum_{k=1}^{M}\frac{k}{rank(r_k)}$$

$$MAP = \frac{1}{|Q|}\sum_{q\in Q} AveP_q$$

ここで，$rank(r_k)$ は正解文書 r_k の順位，$|Q|$ はクエリ総数である。

(2) シーン間検索

通常よく用いられる方法は，予め検索対象をシーン分割し，シーン毎に各単語の TF-IDF 値を並べたベクトルを求めておき，検索質問文から生成された検索ベクトルとの内積が小さいシーンを求める方法である。

シーン間検索の場合には，検索質問文から生成された検索ベクトルの代わりに現在閲覧しているシーンを検索ベクトルにできる．本研究ではこの方法を基準(Baseline)とする。なお，ベクトル間の類似度には Pivoted Document Length Normalization を用いた。

(3) シーンベクトルの線形補間

授業や講演などを検索する場合，検索したい該当シーンに話題の情報が含まれないことがある。Nanjo らは，検索対象を 15utterance unit, 30utterance unit, 60 utterance unit, 全体の重み付き線形和によって検索ベクトルを構成し，良好な結果を得ている。

本研究では，講演等の最初と最後のシーンも検索ベクトルに加える。講演等では，最初に概要を述べることが多い。また講演等の最後にはまとめとして，講演全体の要約が含まれることがある。よって講演等の最初のシーン，最後のシーン，前後のシーンの情報は，話題情報として有用と考えられる．そこで i 番目の講演中の j 番目のシーンの検索ベクトル $V_{i,j}$ を以下のように定義する．

$$V_{i,j} = (1-\beta_s-2\beta_a-\beta_e-\beta_w)T_{i,j} + \beta_s T_{i,1} + \beta_a(T_{i,j-1}+T_{i,j+1}) + \beta_e T_{i,Ni} + \beta_w \frac{1}{N_i}\sum_{k=1}^{N_i} T_{i,k}$$

ここで $T_{i,j}$ は i 番目の講演中の j 番目のシーンの TF-IDF ベクトル，N_i は i 番目の講演のシーン数，$\beta_s, \beta_a, \beta_e, \beta_w$ はそれぞれ最初のシーン，隣接シーン，最後のシーン，講演全体に対応する係数である。このようにシーンベクトルの線形補間を用いたシーン間検索方法を図３に示す。

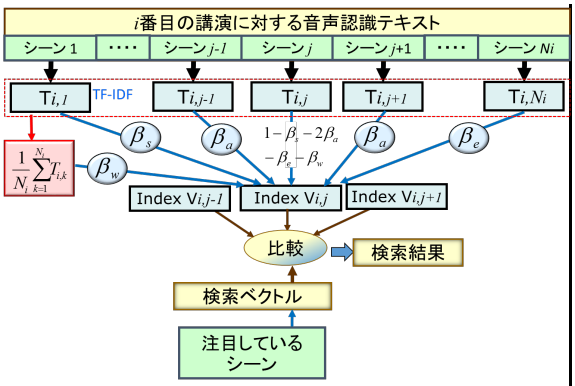

図３　シーンベクトルの線形補間

４．研究成果

(1) 英語の講義ビデオにも対応した関連ビデオ検索システムの開発と公開

① システムの構成

図４にシステムの構成を示す。本システムはサーバ部とクライアント部からなる。

サーバ部では事前に検索対象となるビデオ教材の登録を行う。登録されたビデオ教材の音声部分を抽出し発話区間ごとに分割する。分割された日本語音声は日本語音声認識システム Julius(http://julius.osdn.jp/)によって音声認識される。英語音声の場合は kaldi(http://kaldi-asr.org/)に含まれているツールによって音声認識される。英語の音響モデルおよび言語モデルの学習データには Wall Street Journal を用いる。音声認識結果から、英語の場合は翻訳サイト(http://honyaku.yahoo.co.jp/transtext 等)を利用して日本語に翻訳する。認識翻訳されたテキストからビデオごとの TF および DF を求め保管しておく。

クライアント部(Web ブラウザ)では年度，授業，日付，キーワードを設定して検索する。関連ビデオの検索では、現在閲覧している動画の時刻からシーンを取得し，現在閲覧しているシーンの TF-IDF と検索対象の各シーンの TF-IDF を比較する。TF-IDF ベクトル間の内積が大きいほど閲覧しているシーンとの関連性が高いシーンとして出力する。

②利用方法

年度，授業，日付，検索ワードを指定することでビデオを検索することができる。閲覧したいビデオのボタンをクリックするとそのビデオを閲覧することができる。

図５の動画閲覧ページにおいて「シーン間検索」ボタンを押すと現在閲覧しているシーンを取得し，シーン間検索を行い別の動画から類似するシーンを検索することができる。シーン間検索から別の動画に移動した後、「前の動画」ボタンを押すことで前回閲覧していたビデオの時刻から再生できる。

また，アカウントを登録すればお気に入りとメモの機能が利用できる。ビデオをお気に入りに登録すればすぐにそのビデオを再生できる。メモを登録すればビデオを閲覧した際に前回保存したメモを確認することができる。

③公開 URL

開発した関連講義ビデオ自動収集検索システムを

http://sail.i.ishikawa-nct.ac.jp/whale_new/

に公開した。公開済みの知識を活用したビデオ検索システム (http://sail.i.ishikawa-nct.ac.jp/ whale/)では，一つの教育機関のビデオ検索が可能であるが，本システムの開発により，海外を含めた複数機関のビデオ検索及び関連ビデオシーン検索が可能となった。

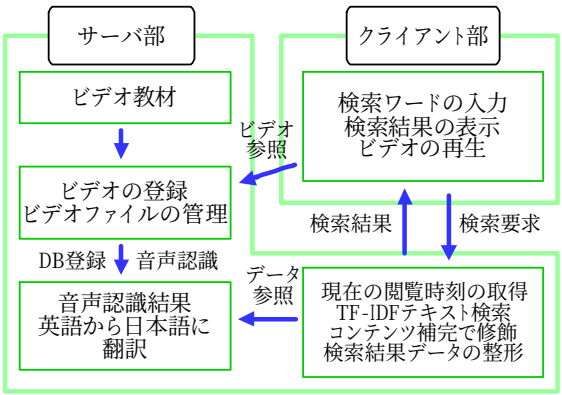

図４　システム構成

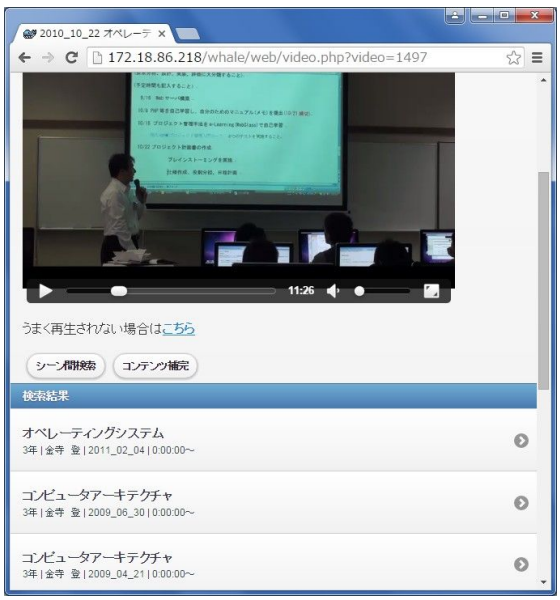

図５ システムのインターフェース

(2) シーン間検索性能の改善

本研究において構築したシーン間音声情報内容検索環境を用いて，現在閲覧している講義ビデオシーンなど注目しているシーンによって関連する音声情報内容検索を行った。

シーンベクトルの線形補間（３．(3)参照）

に用いるパラメータ $\beta_s, \beta_a, \beta_e, \beta_w$ の値は，NITCIR9 を用いて最適化することにより決定された。注目しているシーンをキーとした関連音声内容検索方法の評価には，NITCIR10 を使用した。パラメータ最適値は，最初のシーンに対応する係数 $\beta_s = 0.19$，隣接シーンに対応する係数 $\beta_a = 0.005$，最後のシーンに対応する係数 $\beta_e = 0.005$，講演全体に対応する係数 $\beta_w = 0.11$ であった。これより，最初のシーンおよび講義全体のシーンが話題情報として有用であることが明らかになった。

一般的な質問文による検索結果を図６に示す。隣接シーンの影響を規定する $\beta_a = 0.005$ 及び講演全体のシーンの影響を規定する $\beta_w = 0.18$ を用いた方法は，Nanjo らの方法に相当する。さらに先頭シーンの影響を規定する $\beta_s = 0.19$ 及び最後のシーンの影響を規定する $\beta_e = 0.0005$ を併用することにより，MAP 値が改善された。これらの結果より，先頭シーン及び講演全体のシーンの影響を検索対象シーンに含めることで MAP 値が改善されることがわかる。

シーン間検索結果を図７に示す。シーン間検索においても，質問文による検索結果と同様にシーンベクトルの線形補間の効果が確認できる。図７のシーン間検索結果において，χ^2 検定の結果，シーンベクトルの線形補間は有意水準 1%で統計的に有効であることが確認できた。

なお，シーンベクトルの線形補間の場合，事前に検索対象の各シーンに対して補完処理を行うことにより，検索時に検索時間はほとんど増加しない。

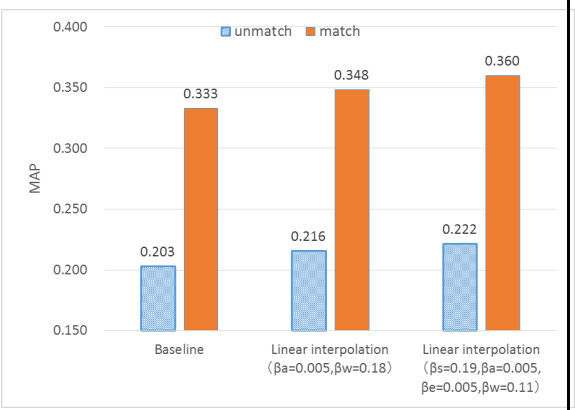

図６　質問文による検索結果

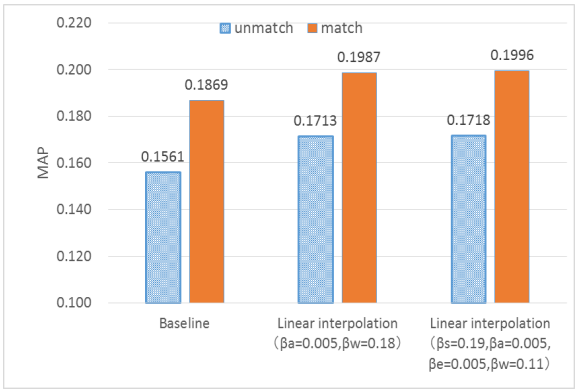

図７　シーン間検索結果

(3) 音声認識性能とシーン間音声情報検索性能の関係

開発した関連講義ビデオシーンを検索できるシステムを利用する上で，どの程度の音声認識性能が必要なのかを知ることは重要である。そこで，シーン間検索性能と音声認識性能の関係を調査するため，書き起こしテキストに誤りを入れたデータに対して検索実験を行った。誤りとして，挿入誤り，脱落誤り，置換誤り それぞれ 0.00～0.40 の範囲で 0.05 刻みの全て組み合わせ(729 通り)に対して調査した。評価尺度には各クエリの平均適合率をクエリ数で平均した MAP(Mean Average Precision)を用いた。

挿入誤り率，脱落誤り率，置換誤り率をそれぞれ P_i, P_d, P_s としたとき，MAP の推定値 $\hat{M}$ を次式で最小二乗近似した。

$$\hat{M} = M_0(1 - \beta_i P_i - \beta_d P_d - \beta_s P_s)$$

ここで，M_0 は誤りがないときの MAP 値であり，係数 $\beta_i, \beta_d, \beta_s$ のシーン間検索時の最適値は，それぞれ 0.195，0.784，0.935 であった。実際の MAP 値と推定された MAP の関係を図８に示す。挿入誤りによる影響が小さいため，図９のように単語正解率でも MAP 値を予測可能と考えられる。

通常の質問文による検索時の係数 $\beta_i, \beta_d, \beta_s$ の最適値は，それぞれ 0.013, 0.418, 0.527 であった。これより質問文による検索も閲覧しているシーンによる検索においても，置換誤り，脱落誤り，挿入誤りの順に検索性への影響が大きいことが確認された。また，閲覧しているシーンによる検索の方が音声認識誤りの影響を受けやすい。閲覧しているシーンによる検索では挿入誤りの影響が質問文による検索に比べて大きいことがわかった。なお，質問文による検索性能が挿入誤りの影響を受けにくいことは，西崎らの結果と一致する。

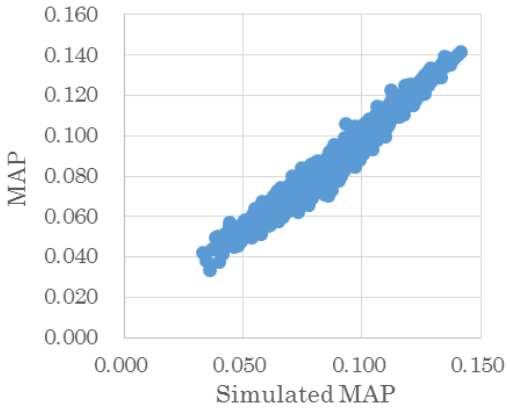

図８　推定された MAP 値

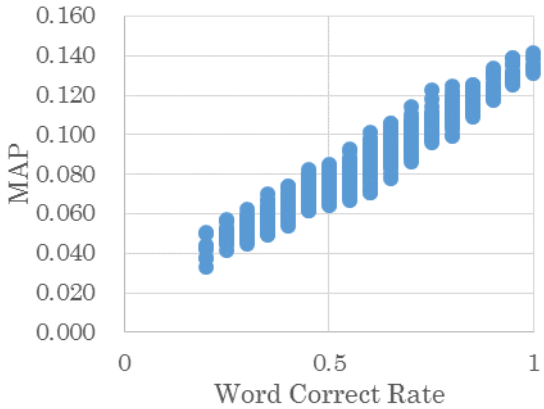

図９　単語正解率と MAP の関係

(4) 国内外における位置づけと今後の展望

収集した世界中の講義ビデオに対して，関連するシーンを検索できるシステム開発し公開した。これまでに現在閲覧しているビデオシーンと関連したシーンを自動検索するシステムの例はない。また，システムのソースコードも公開しているため，拡張が容易である。さらに，学習目的以外の用途でも幅広く応用可能である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

長田忠良，金寺　登，“音声ドキュメント内容検索に関する研究，” 石川高専紀要,48, pp.21-28，2016.3

〔学会発表〕（計７件）

Noboru Kanedera, “Relationship between speech recognition performance and related spoken document retrieval performance,” The 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, 2016.11.

金寺 登，“シーン間音声情報内容検索の性能評価，”日本音響学会 2016 年秋季研究発表会，2016.9.

金寺　登，長田忠良，“閲覧しているシーンによる関連音声情報内容検索，”電子情報通信学会技術研究報告，2015.8.

松井 洸，冨澤幸太郎，東崎 大，東 怜央，金寺 登，“日本語と英語に対応した関連ビデオ検索システムの試作，”平成 27 年度北陸地区学生による研究発表会，2016.3.

金寺 登，長田忠良，“シーン間音声情報内容検索，”日本音響学会 2015 年秋季研究発表会，2015.9.

長田忠良，金寺 登，“音声ドキュメント内容検索に関する研究，” 平成 27 年度電気関係学会北陸支部連合大会，2015.9.

長田忠良，金寺　登，“音声ドキュメント内容検索に関する研究，”北陸地区学生による研究発表会，2015.3.

〔その他〕
ホームページ:
http://sail.i.ishikawa-nct.ac.jp

６．研究組織
(1)研究代表者
金寺　登（KANEDERA, Noboru）
石川工業高等専門学校・電子情報工学科・教授
研究者番号：５０１９４９３１